Markov models of biomolecular systems

By

Robert Edward Arbon

School of Chemistry
University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Science.

July 2021

Word count: 61559

Abstract

Markov models are a popular technique for understanding the dynamics of systems which move through "rough" potentials [1]. In such cases, the system is well approximated as transitioning between discrete states with a set state-to-state probability, independent of its history. Choosing how these states relate to the coordinates of the system (the discretization) and how these are partitioned into metastable sets (the coarse graining) is of central importance to the technique. This thesis contributes to methods for making these choices and applies them to two systems: water diffusion and enzyme dynamics.

Markov models were used to provide an explanation of water diffusion through viscous aerosol particles, where diffusion is known to diverge from typical Stokes-Einstein behaviour. The choice of discretization and coarse-graining techniques came from established methods and heuristics in the Markov modelling literature. The analysis showed that water diffuses by hopping between transient cavities created by the organic fraction of the aerosol particle. For the majority of the time this process is irreversible but the water can also establish local equilibria between clusters of cavities arresting the diffusion process.

A more complex workflow was proposed and evaluated for the case of the aromatic amine dehydrogenase, an enzyme at the heart of the debate surrounding hydrogen tunneling and enzyme dynamics. This workflow used ideas from the statistics and machine learning communities in order to make the modelling process more transparent, efficient and reproducible. The response surface of an MSM - the change in model quality in response to modelling choices - was estimated and optimised using Bayesian optimisation. Statistical model selection techniques for selecting the number of metastable states in a hidden Markov model were evaluated. Theoretical and practical arguments are made in favour of the integrated complete-data likelihood criterion. The benefits of this more elaborate workflow were mixed. The response surface proved useful in creating tests of the sensitivity of inferences to the modelling choices. Many of the modelling choices were shown to not affect the model quality and as a result Bayesian optimisation proved of little benefit. The conformational landscape of aromatic amine dehydrogenase was found to consist of many short lived ($20\,\mathrm{ns}$ to $300\,\mathrm{ns}$) metastable states which slowly interconvert on a timescale of approximately $1.2\,\mu\mathrm{s}$. However, the simulations had moved away from their reactive conformations and so the implications for understanding reactivity were limited. In addition, these results could not be validated and sensitivity tests cast doubt on the robustness of this conclusion. The source of these problems was investigated and several solutions were proposed.

DEDICATION AND ACKNOWLEDGEMENTS

This thesis is dedicated to the community of scientists and engineers of all kinds which have made this thesis possible. In particular, however, I would like to acknowledge and thank the following people.

First, Toni Mey, who has provided so much detailed and substantive feedback on my work, promoted my cause, and been a constant source of encouragement throughout my PhD: from correcting my coding errors on the fly in my first group talk, to last minute analyses in this thesis. To Adrian Mulholland for reading and critiquing this thesis and for support and encouragement throughout my PhD. To Eric Lang for his feedback and useful discussions on chapter 6 and to my good friends Fred Cascarini and Emily Lythell for holding my hand through the submission process and for constant emotional support throughout my time in Bristol.

I would also like to thank the staff at the Jean Golding Institute for providing such a friendly, encouraging and stimulating place to work, in particular: Patty Holly, Natalie Thurlby, Kate-Robson Brown, Rachel Prior and Liz Green. And to the Unite House family of 2018, Sandy Payne, Cait Houlbrook, Felipe Côrte Real De Camargo and the others: it was emotional.

Completing this thesis would not have been possibly without the support of my friends and colleagues: Attila Jantsek, Balázs Hornung, Mélangi Panagi George Holloway, Lisa May Thomas, Alex Jones, Tom Mitchell, Mark Wonnacott, Silvia Amabilino, Mike O'Connor, Rob Kent, James Hall, Helen Murphy, Zara Cassid, Elias & Rose & family, Gareth & Louise & family, Tom Oliver, and to all those in the Muholland and Glowacki groups. Special thanks must also go to CJ the Wondrous and their partner James for being such amazing friends and providing a wonderful home for me while I write. Lastly my family: my uncle David, cousin Christine, my sister Katie, her husband Mike and of course my mother Susan whose constant, unconditional love and support sustained me throughout this process.

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE: .....

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

This thesis describes the use of statistical model selection and optimisation techniques applied to Markov models (MM) for describing the slow conformational dynamics of two biomolecular systems: water diffusing through a sucrose matrix and the conformational landscape of aromatic amine dehdyrogenase (AADH).

1.1 The importance of conformational changes in biochemical systems

Quantitatively describing the conformational changes in biomolecular systems is of central importance for understanding their function, chemical and biological properties. Conformational changes are at the heart of enzyme catalysis [2–5]. For example, triosephosphate isomerase (TIM) catalyses the isomerisation dihydroxyacetone phosphate and (R)-glyceraldehyde-3-phosphate [6] and is considered a paradigmatic system for enzyme catalysis [7], having been studied by molecular simulation since at least 1987 [8, 9]. The catalytic process requires the closure of a loop (through approximately 7 Å) after ligand binding which creates the necessary conditions to allow the isomerisation to occur, opening again to allow product release [6]. More extensive conformational changes are present in other systems. Dihydrofolate reductase (DHFR) catalyses the reduction of 7,8-dihydrofolate to 5,6,7,8-tetrahydrofolate [10] with at least five kinetically distinct intermediates detected across the catalytic cycle [11]. The Met20 loop adopts at least three distinct conformations across the catalytic cycle [12] with mutation experiments demonstrating its importance in the rate determining step [13]. The importance of conformational change in the relative positions of DHFR's two sub-domains has also been implied in transition state stabilisation and product release [12]. Conformational changes have also been invoked to understand non-linear Arrhenius plots (which relate the rate of reaction to its activate barrier) [14, 15]. The cold adapted psychrophilic α-amylase enzyme has an optimum rate for catalysis well below its melting temperature [16]. Macromolecular rate theory [17]

posits changing heat capacities as an explanation, however, simulation studies found an equilibrium between a reactive and an unreactive enzyme-substrate complex a more fitting explanation [16]. There is evidence that population shifts in their conformational landscapes are the mechanism by which enzymes evolve to catalyse different substrates and reactions [18–20]. Other biochemical examples of conformational change include G protein-coupled receptors (GPCRs), a large family of transmembrane proteins involved in extracellular sensing and signalling which are responsible for olfaction, vision and taste [21]. They transmit information from outside of the cell by way of ligand binding which induces a series of conformational changes which in turn triggers a response within the cell [22]. Conformational changes are also central to allosteric modulation - where binding of a ligand at one protein site affects the function of the protein at a separate site [23]. A classic example is of the enzyme aspartate transcarbamoylase which binds asparate and carbamoyl phosphate at the active site and catalyses their reaction into carbamoyl-aspartate. Binding of cytidine triphosphate, a product formed downstream in the metabolic chain, inhibits the reaction by binding to a site $60\,\text{Å}$ from the active site [24]. Upon binding, conformational changes occur which shift the enzyme to the unreactive conformational state from the reactive site. Adenosine triphosphate binds similarly but instead shifts the equilibrium in the opposite direction, activating the reaction. Allostery thus allows metabolic pathways to react to the cellular environment. The function of GPCRs and other protein receptors is also guided by allosteric mechanisms [25]. Finally, large conformational changes are also implicated in the mechanism by which proteins associate with specific parts of DNA, thus enabling a whole host of cellular functions from gene regulation to DNA replication [26].

1.2 Learning conformational dynamics from molecular simulations

Computational approaches to studying conformational dynamics are important because they allow descriptions with high temporal and spatial resolution beyond the scope of most experimental techniques [27]. A popular approach is to evolve the systems' equations of motion using molecular dynamics (MD) to produce a set of trajectories through phase space. These trajectories can be used to reveal conformational transition pathways, metastable states and other properties of interest by estimating the relevant regions of the free energy landscape (the free energy with respect to a given set of coordinates) [28].

There are a wide range of techniques for understanding conformational dynamics from MD simulations, the suitability of which depend on current knowledge of the system and the questions being asked. Path based techniques such as transition path sampling (TPS) [29–31], transition interface sampling (TIS) [32] and forward flux sampling (FFS) [33] all start with two specified metastable conformations, A & B, and can be used to estimate rate constants and reaction coordinates of the reaction A $\rightleftharpoons$ B without previous knowledge of reaction pathways. TPS works by first proposing a reaction pathway between A and B. Then, a statistical ensemble of pathways is generated from this initial path using a stochastic algorithm which takes into account the potential

energy of the system (Metropolis Monte Carlo). TIS and FFS are similar to TPS but define interfaces which separate A and B based on an order parameter[1]. Molecular dynamics simulations are then used to estimate the flux between the interfaces and hence the transition rate between A & B (FFS relaxes the assumption of equilibrium dynamics required in TIS and TPS). These techniques have been applied [34–36] to the model protein folding system Trp-cage [37] to elucidate its folding pathway and have replicated some of the experimental microscopic rates in the folding pathway. String methods are similar but posit a discrete set of states along a path (or string) of fixed length and moves these states such that the string corresponds to minimum free energy path between A and B [38, 39]. String methods have been used to determine, for example, the mechanism and binding free energy of platinum based drugs to DNA [40].

Directional milestoning [41–43] is a technique which does not require knowledge of specific metastable states, only a collective variable known to be related to the important conformational changes. First a set of states which cover the relevant conformational space, known as anchors, are selected. The milestones are sets of conformations which separate (as measured by the collective variable) the anchors from one another, and are used to restart trajectories, calculate reaction coordinates and kinetics rate constants. Milestoning has been used to understand the selectivity of DNA reverse transcriptase [44] and the mechanism and rate of tryptophan permeation through cell membranes [45].

When no previous information such as known metastable states or appropriate collective variables are known, more abstract statistical and machine learning methods have been increasingly shown to be important. Principal component analysis (PCA) finds the linear combinations of features of a molecule, such as the α-carbon coordinates of a protein, which explain the variance observed in a trajectory (the principal components are mutually orthogonal and explain decreasing amounts of the variance) [46]. Molecular motion can then be approximately described by a handful of principal components, rather than the full set of atomic coordinates. In reference [47] the authors used PCA of the heavy atom coordinates of the ribonuclease barnase to identify the highly flexible (high variance) regions of the enzyme which they related to the enzyme's activity and stability. The authors of reference [48] used PCA in the protein lysozyme to identify highly flexible regions which are related to the opening and closing of the active site (the "essential" degrees of freedom in their language). One of the draw backs of PCA is that only linear combinations of features make up the principal components. Kernel PCA, which incorporates non-linear transformations of input features, has also been developed and used with TPS simulations to extract a reaction coordinate for the reaction of lactate dehydrogenase [49, 50]. Multidimensional scaling (MDS) [51] is similar to PCA in that it seeks to represent high dimensional data with a smaller number of combinations of input features. Rather than finding components to capture the variance, MDS preserves distances between observations. MDS has been used to characterise the conformational states and track simulation convergence of bovine pancreatic trypsin inhibitor [52]. Similar machine learning methods, which

[1] a quantity such as the root mean square deviation, which, while not a true reaction coordinate, varies between the two states.

find low dimensional representations of the dynamics while preserving various metrics, have also been used. Isomap [53] preserves the distances between conformations on a curved surface (manifold), the geometry of which is inferred from the observed conformations. A computationally efficient method of Isomap (SciMAP) was used to determine the protein folding reaction coordinate for SH3 domain [54]. Sketch-map [55] preserves only certain subsets of distances deemed to be important and has been used to understand the unfolding dynamics and the effect of point mutations of a beta hairpin polypeptide [56]. Diffusion map [57] and locally scaled diffusion map, preserve diffusion distances (i.e., how easily states can diffuse to one another). They have been used to characterise folding pathways in a number of small proteins: Trp-cage [58], a beta-hairpin [59] and in Microcin J25 [57].

1.3 Markov models and their applications

An alternative to the techniques described above are Markov models. Markov models provide a framework for classifying conformations into metastable states, finding reaction pathways and estimating kinetic and thermodynamic quantities. While they are able to incorporate knowledge of important order parameters or features, they do not require such knowledge [60, 61].

The central idea [62] is that for complex systems, over a sufficiently long periods of time, the rate at which the system transitions out of region A and into region B of configurational phase space is not dependent on how the system arrived at A. In other words, these transitions are "memoryless". In mathematical notation this is [63]:

$$\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = \mathbf{p}(t)\mathbf{K}, \tag{1.1}$$

where the i'th element of $\mathbf{p}$ represents the probability of the system being in a region of phase space labelled by i; K_{ij} is the rate of transitioning from region i to j, and $K_{ii} = -\sum_{j\neq i} K_{ij}$. In addition to being "memoryless", equation 1.1 also implies that the rate matrix $\mathbf{K}$ does not change with time, i.e., the system is "stationary" [62]. The solutions to this equation describe how the probability of the system being in discrete regions of phase space changes smoothly over time. The justifying assumption for memoryless transitions is that biomolecular systems have a free energy surface (the free energy with respect to some set of coordinates) which is is characterised by many local minima, arising due to the many degrees of freedom afforded by its large structure (e.g., rotations about bonds or dihedral angles). The system resides in these minima for a sufficiently long period of time that transitions between them become independent of one another. While this is not the case for very short timescales, over a sufficiently long time τ, the Markov lag time, this becomes a valid assumption. Memoryless dynamics over this timescale will also be referred to as "Markovian" throughout this work.

The Markov model approach to solving equation 1.1 is to consider a discrete time process ($t = k\tau,\ k = 1,2,\ldots$) and to partition the configurational phase space into n discrete states so that the dynamics of the system can be described by an $n \times n$ transition matrix $\mathbf{T}$ [64]. Each element of

$\mathbf{T}$ describes the conditional probability of the system jumping between states over the Markov lag time, τ i.e., [63]:

$$\mathbf{p}((k+1)\tau) = \mathbf{p}(k\tau)\mathbf{T} \tag{1.2}$$

The eigenvectors and eigenvalues of $\mathbf{T}$ represent the associated slow dynamic processes, such as protein folding or loop opening and closing, and their associated timescales [64].

Applications of Markov models (MMs) are concentrated on biomolecular systems and form an intrinsic part of the biomolecular simulation tool-box [27]. Applications include modelling both protein folding pathways [65, 66] as well as intrinsically discorded proteins [67]. MMs have been applied to enzyme systems and used to elucidate, for example, ligand docking pathways [68] and regioselectivity mechanisms in cytochrome p450 [69], the conformational heterogeneity in the important cancer target SETD8 [70], loop dynamics in triosephosphate isomerase [6], and allosteric effects in cyclophilin A [71]. Other applications include self-assembly [72] and dimer formation [73] of amyloid peptides, identifying important conformations in drug targets to improve drug docking free energy calculations [74], and rational drug design [75]. There has been comparatively fewer applications of MMs on smaller systems (whose kinetics tend to be derived from quantum mechanical and thermodynamic data [76, 77], rather than statistically estimated from MD data), however, one recent example used MMs to determine hydrogen bond rearrangement in liquid water [78].

Early MM construction consisted of describing the conformational dynamics of systems in thermal equilibrium by constructing only a handful of discrete states, and modelling the dynamics as Markov chain also known as a *Markov state model* (MSM). For example the authors of reference [79] investigated the folding of a heptapeptide into a β-hairpin conformation in a solution of methanol. To decide whether the folding process is a memoryless process (i.e., conforms to equation 1.1) they estimated a four state MSM and compared the transition probabilities implied by the model (the elements of the 4×4 matrix $\mathbf{T}$) to those observed in the MD trajectory. The four states were based on their geometric similarity in the space of a principal component analysis of the peptide backbone coordinates. They found them to be in good agreement and so concluded that their reduced dimension description of the folding process was valid.

However, the more common approach [60, 80] is a two stage process. In the first stage, frames from MD simulations are geometrically clustered into n discrete *microstates* (where typically $n \lesssim 1000$) and the elements of $\mathbf{T}$ are estimated in this microstate basis. The purpose of this discretization is to allow an precise description of the eigenvectors of $\mathbf{T}$ in terms of these microstates, the eigenvectors in turn describe the various conformational transitions [81]. The assumption behind the validity of this approach is that with a fine-grained definition of microstates (i.e., each microstate is structurally very similar) their geometric similarity is enough to guarantee their kinetic similarity. In other words, if a set of MD frames are clustered into the same state i (the elements of $\mathbf{p}$ in equation 1.1) then they are all accurately described by the same set of rates to other discrete states ($K_{i,j}$). Care must be taken here as geometric similarity does not always imply kinetic similarity, so that structures which appear similar according to some metric may have very dissimilar kinetic properties

[63, 82–84]. The methods for creating this fine-grained kinetic model as well as alternatives to geometric clustering will be described in this section.

The second stage makes use of the fact that very often there is a separation in timescales between the slow processes of interest and other processes. To take the triosephosphate isomerase example of earlier, loop 6 opens and closes on the timescale of $100\,\mu s$ [6], whereas the microstates might resolve oscillations of contact distances or side chain rotations (which maybe of the order $1\,ns$ to $100\,ns$) which are not relevant to the overall loop motion. This fact allows coarse-graining these microstates into a handful, g, of *macrostates* based on their kinetic properties. The macrostates are usually defined such that microstates have a low probability of transitioning between the macrostates, compared to inter-conversion within a macrostate [64, 66, 85]. Coarse-graining will be discussed in depth in sections 1.7 and 1.8.

One approach to creating accurate MSMs is to focus on finding the "essential degrees of freedom" of the system [62, 85] i.e., a small number of features (compared to the number of atomic coordinates) which describe the slowest conformational processes in the system. This is justified when the slowest conformational changes are the ones of interest. Examples of features for describing protein folding include the root mean square deviation from the crystal structure, the fraction of contacts found in the crystal structure, or even thermodynamic quantities like energies arising from solvent interactions [86]. There are several benefits to identifying these features before clustering atomic configurations into microstates to estimate an MSM. First, geometric similarity as measured in the space of the features is more likely to correlate with kinetic similarity than atomic coordinates. Second, it reduces the computational effort required to cluster MD frames into microstates. K-means clustering, a popular method for performing geometric clustering, has a computational complexity which scales with the number dimensions [87]. So reducing the number of variables from 1000–10 000 (the order of the number of atomic coordinates for a typical protein[2]) to 10–100 for typical number of molecular features, e.g., dihedral angles of the protein backbone, represents a large reduction in computational complexity.

As already discussed there have been many machine learning techniques for reducing the dimension MD trajectories, e.g., PCA, multidimensional scaling, iso-map, and sketch-map. However, these techniques do not directly address capturing *slow* dynamics. For example, the problem with using PCA as a preprocessing step before clustering into microstates is that the principal components explain the greatest *configurational* variance [81]. To address this drawback, time-lagged independent component analysis, TICA [81, 89], was introduced whose components capture the greatest *kinetic* variance. The total kinetic variance describes the ability of the TICA components (or any basis set) to capture the slow dynamics of the system [90]. TICA identifies linear combinations of the atomic positions which are maximally correlated at a given lag-time (also referred to as τ but not necessarily the same as the Markov lag-time). TICA is a stand-alone technique whose eigenvectors are the optimal linear approximation to the eigenvectors of $\mathbf{T}$ [91]. TICA can also be used a preprocessing

[2]The modal number of atoms of structures in the Protein Data Bank is 2000–3000 according [88]

step to reduce the dimensionality of MD data prior to estimating a MSM by projecting only onto the first m TICA components (where m has to be specified). Using TICA as a preprocessing step in MSM construction has been systematically investigated and shown to be more accurate at capturing the slow dynamics than both PCA and no preprocessing at all [92]. In this work the authors randomly chose different MSM *hyperparameters* - different modelling choices such: as the choice of protein feature, dimensionality reduction using PCA, TICA (or none at all), the TICA lag time and number of retained TICA components, and the number of microstates. Using these hyperparameters they estimated MSMs and evaluated their accuracy in capturing the the folding dynamics of twelve fast-folding proteins [93] (using the GMRQ score, described below). They found that, on average, models built with TICA had significantly greater accuracy. However, using TICA brings with it two new modelling choices: the TICA lag-time τ and the number TICA components, m, onto which the protein features are projected. These TICA lag time affects what slow processes are captured by the TICA projection, while the number of retained components affects the accuracy of the description. The effects of these parameters will be investigated in chapter 6.

The variational approach to conformational dynamics, (VAC [91]), cast estimating MSMs and TICA as a variational optimisation problem. VAC showed that TICA and MSMs were the optimal description of the slow dynamics for a *given* continuous (for TICA) or discrete (for MSMs) basis set. The authors of reference [94] extended this idea and showed that the same variational principle could optimize the basis sets themselves. The key innovation of this work was to combine cross-validation [95] and the variational principle to score a given basis set using the generalized matrix Rayleigh coefficient, GMRQ [3]. Cross-validation is a technique for estimating how well models perform on unseen data - i.e., it estimates the model generalizability. MSM theory for biomolecular simulations was then broadened with Koopman models to encompass simulations of systems out of thermal equilibrium [96]. With the variational approach to Markov processes (VAMP [97]) the theory of MSMs and Koopman models was unified into one conceptual framework. This increased the scope of MMs and presented a range of model scoring metrics, called VAMP scores, of which the GMRQ was a special case. These theoretical advances have allowed the development the following iterative optimisation MM pipeline, starting with a set of MD trajectories:

1. project atomic coordinates on to important features;
2. project features onto TICA components ("essential degrees of freedom");
3. geometrically cluster MD frames into discrete microstates;
4. estimate an MSM and score using a VAMP score;
5. repeat the previous steps by varying the type of feature, number of discrete states, etc., until a satisfactory VAMP score has been achieved.

Other approaches to building MSMs exist which do not focus on find the essential degrees of freedom as a preprocessing step. VAMPnets [98] still utilises the variational framework but instead of

[3] similar to finding variationally optimised electronic wave function basis sets

discretising MD trajectories it uses a deep neural network to learn continuous, non-linear estimates of the eigenvectors of $\mathbf{T}$. It uses the atomic coordinates of MD trajectories directly, essentially condensing the whole pipeline into a single process. Enspara [99] is a package which facilitates clustering large volumes of MD data without the need to perform dimensionality reduction first. By focusing on producing fine-grained microstates without the need to project the coordinates onto the slowest collective variables, this technique retains a larger range of dynamic processes (not just the slowest ones). For example, coordinate trajectories of a cefotaximase [100] were clustered into fine grained microstates based on similar values of RMSD relative to a crystal structure. As a result, the authors were were able to reveal important, rapidly interconverting, conformations, not captured in X-ray and other structural data, to explain the enzyme's specificity and antibiotic resistance.

Another similar approach to Markov state models for understanding conformational dynamics is discrete path sampling (DPS [101]). DPS solves equation 1.1 by creating microstates based on their kinetic properties rather than their geometric properties, using the potential energy surface rather than MD trajectories. First, a database of local minima (which define the discrete microstates) and saddle points (corresponding to transition states between the microstates) are created by geometry optimisation of the potential energy surface of the system. Then, the elements of $\mathbf{K}$ can be estimated using transition state theory [102]. This method is limited by the number of degrees of freedom (which increase the fluctuations in the potential energy surface) which for biomolecules can become prohibitively large [63]. However, using implicit solvent models to limit the number of degrees of freedom, the conformational dynamics of small and medium sized systems have been investigated. These include the folding dynamics of met-enkephalin [103] and tryptophan zipper peptide [104], characterising the free energy surface of intrinsically disordered proteins [105], and the effect of point mutations on the a coiled-coil peptide [106].

Chapter 2 sets out the theory of MMs relevant to this thesis which focuses on MSM estimation, TICA for preprocessing, and variational optimisation of basis sets using VAMP scores.

1.4 Simplified Markov state model construction

The MSM literature has concentrated on large biomolecules because their hierarchy of atomic motions [107] give rise to a rugged, free energy surface with "memoryless" conformational transitions. As already mentioned, the key to success of the MM process is the creation of a good set of microstates to represent the dynamics of the system. When the focus is on the slow dynamics, identification of relevant features is key to this process. For biomolecules the process is complicated by the large number of potentially relevant features and other modelling choices which cannot be determined *a priori* [108]. Instead, the iterative optimisation process delineated in the previous section, is used to choose the best set of modelling choices, however, this can be computationally intensive. In contrast, for systems with a much smaller number of relevant degrees of freedom, chemical intuition and visualisation techniques can be used to guide the choice of collective variable.

Chapter 3 describes computational and experimental work designed to understand the diffusion of a single water molecule through a sucrose matrix, designed to mimic the conditions of water diffusion in secondary organic aerosol (SOA) droplets [109] (i.e., aerosol consisting of organic molecules dissolved in water [110, chapter 1]).

The motivation for studying aerosols in general is that they have wide ranging impacts on human and planetary health [111], from smog in cities [112], directly affecting the radiative balance of the atmosphere by altering its chemical composition [113], and indirectly through their effect on cloud formation [114]. As to SOAs in particular, they have been increasing recognised as an important source of total atmospheric aerosol, alongside the more well known primary sources such as ocean spray, smoke from natural and man-made sources [115]. The water content of SOA influences its chemical reactivity [116] and physical properties like size and refractive index [117, 118]. Predicting water diffusion in SOA is therefore important to explaining a range of SOA phenomena.

The Stokes-Einstein (S-E) definition of diffusion, D, relates the viscosity of the solvent, η, to the hydrodynamic radius of the solute, a, at a temperature T is given by [119, chapter 17]:

$$D = \frac{k_{\mathrm{B}} T}{C \pi \eta a}, \tag{1.3}$$

(C and k_B are constants). For SOA droplets existing in the low humidity parts of the atmosphere, water evaporates to the point that the organic constituents of the particle become the dominant mole fraction leaving water as the solute [120–122]. In these regimes large deviations from S-E diffusion occur [120, 123, 124]. There is a continuing debate over the applicability of the S-E description of diffusion in SOA droplets (see chapter 7 of reference [111] for a review), with different *ad hoc* modifications of the S-E being suggested [125–127] as well as a case being made for entirely new explanations [128]. For the system studied in this thesis, the observed diffusion rate is much larger than that predicted from the observed viscosity of the sugar component and the water radius using the S-E equation. These deviations occur when the viscosity is so high that the aerosol droplets start to transition to a glassy state [129]. The motion of sucrose matrix in this case becomes slow on the timescale of the motion diffusing water molecules, but not so slow that it can be considered stationary.

The aim of chapter 3 is to both add to the debate over water diffusion in SOA by suggesting a microscopic mechanism for water diffusion in a system with large deviations from S-E behaviour, and to show that the iterative, variational approach to building Markov state models described in the previous section is not always necessary. Instead a simplified approach utilising chemical knowledge and intuition can be used to construct valid and informative MSMs. The MSM approach is justified because the interactions of the water molecule with the much large sucrose molecules creates a sufficiently complex free energy landscape that the "memoryless" assumption for configurational transitions holds. However, the assumption that the transition rates do not change with time was not met due to the slow but persistent motion of the sucrose matrix. Another aim of this chapter was therefore to demonstrate a simple way of accounting for non-stationary transition rates when constructing MSMs.

1.5 Evaluating Markov state model performance

The MM analysis pipeline described so far, consists of first transforming MD trajectories into features (the essential degrees of freedom, χ), then reducing the dimension with TICA, discretizing the TICA components into n microstates, and finally estimating the MSM. The modelling choices or *hyperparameters*, (χ, τ, m, n), create the MSM basis set, which in turn determine the accuracy of the resulting MSM, and so a method of evaluating the performance of these hyperparameters is needed. While the ground truth of the kinetic processes is not available, the initial way forward came through cross-validation and the GMRQ [94].

The innovation in reference [94] was to create a model score, the GMRQ (the Rayleigh trace from quantum mechanics), which could be used to judge the quality of the model choices while accounting for the tendency of models to fit to noisy signals in the data (over-fitting). This was achieved through cross-validation [95]: a model is estimated using a portion of the data and scored on the remaining data. Maximizing the cross-validated GMRQ by varying the hyperparameters increases the accuracy of the eigenvectors [94]. The GMRQ is a special case of the first VAMP score, VAMP-1, while maximizing the total kinetic variance is the same as maximizing the VAMP-2 score. These VAMP metrics completed the analysis pipeline [130] which now can be summarised as: i) transform MD trajectories into features, χ, ii) select reasonable choices of hyperparameters (features, TICA parameters, number of discrete states) and calculate the cross-validated VAMP-2 score, iii) change the hyperparameters and repeat analysis, iv) stop when the VAMP-2 score stops increasing.

1.6 Hyperparameter optimisation

Choosing the hyperparameters which maximize the VAMP-2 score is a 'black-box' optimisation problem [131], so called because no gradient information on the response of the VAMP to the hyperparameters is available. This is a common problem in the machine learning community where models have many parameters and may take days to train [132]. In this case it is not feasible to exhaustively search through combinations of hyperparameters. A popular method for optimising large sets of hyperparameters is Bayesian optimisation (also known as sequential model based optimisation, SMBO) [133–136]. The idea behind Bayesian optimisation is that there is an objective function which is costly to optimise [137, 138] (in this case the VAMP-2 score). So instead of optimising this directly, the BO procedure builds an statistical model of objective function known as a *surrogate function* or *response surface*, using randomly sampled values of the objective function. Having built an initial response surface, searching for the next hyperparameter to evaluate is guided by an *acquisition function*. These can be selected or adjusted to trade off high-uncertainty regions (the 'explore' regime) of the response surface with the high-value, low-uncertainty regions (the 'exploit' regime) [138]. A suggestion is evaluated, the response surface updated and the process repeats. Bayesian optimisation for hyperparameter optimisation is popular, as the number of packages designed

for this purpose will attest (this list is non-exhaustive): Hyperopt [139]; sequential model-based algorithm configuration [133], SMAC; BayesOpt [140]; Spearmint [134, 141–144], GPyOpt [145], DragonFly [146]; Auptimiser [147]; and Osprey [148]. A popular choice of response surface model is a Gaussian process (GP) [149], a highly flexible type of model which fits naturally within the Bayesian optimisation paradigm [138]. Indeed, six of the eight packages listed here all implement some kind of Gaussian process as their response surface model.

The aim of **chapter 4** is to demonstrate the use Bayesian optimisation to optimise the MSM hyperparameters using cross-validated VAMP-2 scores of the model system alanine dipeptide. In addition, the parameters of GPs are explored as a way to describe the relevance of hyperparameters in determining the VAMP-2 score. This chapter lays the ground-work for performing a similar analysis on AADH in chapter 6, in particular: how to fit and interpret GPs and how to use GPs with Bayesian optimisation to optimise hyperparameters.

1.7 Coarse-graining

The first stage of the Markov modelling processes is to create n microstates via an optimal set of MSM hyperparameters by maximising the kinetic variance. The second stage in the process is to coarse-grain potentially thousands of microstates into a handful of macrostates to create a more interpretable model.

However, coarse-graining an existing MSM is not the only approach to gaining insight into the conformational landscape of biomolecules. There are other statistical clustering techniques that have been used for this purpose. The authors of [52] used hierarchical clustering [150, chapter 10] to group MD frames into groups with mutual root mean square deviation (RMSD) in their alpha-carbon positions below some small threshold value. Hierarchical clustering shows how conformations cluster together as the threshold RMSD is increased. In this way the conformational landscape at different levels of spatial resolution can be determined and the number of clusters determined by other criteria. In reference [52] the number of clusters was chosen so that members of each cluster were in the same potential energy minima (albeit this was imperfect as the clustering was still based on geometric similarity, see discussion in section 1.3). In reference [151] the authors used a neural network clustering algorithm, ART-2′ [152], to investigate the folding mechanism of a pentapeptide. Folding events were described by up to six different clusters where the clustering took place in the space of residue dihedral angles. The number of clusters was determined by considering the size of the clusters in the dihedral space, as opposed to considerations arising from observed conformational changes.

The main drawback of clustering based on geometric measures of similarity are that metastable macrostates are actually defined by their kinetic properties (i.e., conformations in a metastable macrostate undergo rapid, mutual, inter-conversion over a given timescale) which are not necessarily the same [85] as configurational similarity, as already discussed in section 1.3. Kinetic clustering dates

back to at least 1969 when Kuo and Wei [153, 154] investigated the conditions under which both exact and approximate coarse-graining of systems of coupled first order reactions could occur. The term *exact* implying that the coarse-grained description gave rise to a kinetic description consistent with the underlying microscopic kinetics. Hummer and Szabo [155] tackled the problem of how to define an appropriate coarse-grained rate matrix for a given coarse-graining scheme. i.e., given a mapping of micro- to macrostates, what is the most appropriate way of defining the rate matrix? They derived expressions for coarse rate matrices which are exact for non-Markovian dynamics (i.e., for systems where transition probabilities are dependent on the history of states visited). They also derived expressions for the case of Markovian dynamics which, while approximate, ensured that the cross-relaxation times between macrostates were consistent with those in the microstate picture. This was later shown to be equivalent to ensuring the mean-first passage times were preserved between the macrostate and microstate pictures [156]. This work has been used as the basis of a method for identifying not only metastable macrostates, but also the comparatively short lived transition states [157], which are known to increase the accuracy of the Markov model description [64].

Other methods, solely based on identifying metastable macrostates have been developed. The underlying idea behind these methods is similar i.e., that kinetic properties of the microstates should determine the assignment of micro- to macrostates, however they differ in how this is achieved. Perron Cluster Cluster Analysis (PCCA) [158] and its subsequent 'robust' alternative PCCA+ [159] were proposed in 2000 and 2005 respectively and were the first methods to explicitly coarse-grain MSMs. They use the properties of the slow eigenvectors of $\mathbf{T}$ to group microstates and are still being utilised, see for example the recent references [6, 160, 161]. Many other methods have been proposed [162–167], some of which have been quantitatively compared in reference [168] and all of which will be described in more detail in chapter 5. However, one of the most popular methods are hidden Markov models (HMMs) [169] which are dynamical models in which the microstate/macrostate coarse-grained structure is directly incorporated into the model definition is estimated from the data. HMMs are models of a Markovian process between g *hidden* macrostates i.e., states which are not directly observed in the data. While in a macrostate the system emits randomly, according to a probability distribution, to one of a set of *observed* microstates, which *are* seen in the data. The hidden states correspond to metastable macrostates and the emission distributions define the mapping between the macrostates and microstates. HMMs are well studied [170, 171] and have been used in many different areas of science from speech recognition [171] to animal movement [172] and have been applied to a large number of biochemical systems [173–186].

1.8 Number of metastable states

When coarse-graining MSMs (or performing any type of cluster analysis) a key parameter is the number of clusters g, e.g., does the data support the hypothesis of $g = 2$ or $g = 3$ (say) clusters

[187]. Choosing the value of g (or any parameter not estimated from the data) is known as model selection [150, chapter 7]. Choosing the optimum value of g is important as each macrostate pertains to conformations important to the dynamical process being studied [107]. If the number of macrostates modelled are too few, then important conformations will be lost, whereas with too many macrostates, the model loses its interpretability and can potentially create macrostates which are artifacts of noise in the data, a processes called 'over-fitting' [150, chapter 7]. The dynamics of proteins are hierarchical [107] with short lived states aggregating to longer lived states, and as such kinetic clustering must always be in relation to some timescale. However, even given this timescale coarse-graining methods do not automatically select the number of macrostates. A general approach to determining the appropriate number of macrostates is to look for gaps in the eigenvalues of the transition matrix or its implied timescales [64, 94, 158]. The number of slow process defined by this gap defines the number of metastable states. However, due to poor microstate construction or insufficient sampling, identifying a clear cut gap is not always possible [168]. In addition, this also does not allow for easy identification of macrostates corresponding to transition-states [157]. A general method for MSMs using Bayesian statistics has been developed, which takes as its data the mapping between the micro- and macrostates [188] and so is independent of clustering method. To decide on an appropriate number of macrostates, the Bayes factor (the Bayesian weight of evidence for a particular hypothesis [189]), for different numbers of macrostates is calculated and used to select g. The evidence is proportional to the probability of observing the microstates *given* the particular coarse-graining and data [188]. This method is versatile and naturally takes into account model over-fitting [188] but is computationally intensive.

Hidden Markov models are distinct from the other techniques in that they can be estimated by maximizing a likelihood function [167, 169] i.e., the probability of observing model parameters given a set of data. Maximum likelihood models have a wide range of model selection techniques available to them which are not explicitly related to Markov processes but are nevertheless applicable because the Markov property is subsumed into the likelihood function [150, chapter 7][187, 190]. Some popular techniques include cross-validation [95], the Akaike information criterion (AIC) [191], the Bayesian information criterion (BIC) [192], and cross-validation of the log-likelihood (CVLL) which have all been used to estimate the number of macrostates in HMMs [193]. The AIC uses the likelihood to approximate the out-of-sample predictive accuracy of the model, whereas the difference in BICs for two models is approximately equal to the Bayes factor for those models (this is directly related to the Bayes factor approach of [188] described previous). Both the AIC and BIC benefit from requiring negligible extra calculation once a model has been estimated and have additionally been used to select the number of microstates in MSMs of conformational dynamics [194] as well as being ubiquitous for general model selection [150, chapter 7]. A BIC-like criterion called the integrated complete data likelihood (ICL) [195] has been derived specifically for clustering methods such as HMMs and mixture models (which group observations into macrostates, albeit without Markov dynamics) [190]. The ICL differs from the BIC and Bayes factor approaches in

that the evidence it considers is proportional to the probability of observing the microstates *and* the coarse-graining *given* the data [190, 195]. The ICL has been used extensively [190] for mixture models and a recent assessment [196] finds it performs well across a range of types of mixtures. The CVLL, BIC, AIC and ICL, have been utilised to determine the number of macrostates in HMMs, but yet not within biomolecular dynamics context.

Chapter 5 explores the utility of approximations to the Bayes factor and similar criteria for determining the optimal value of g: the Bayesian information criterion, BIC [192], the integrated complete data likelihood criterion, ICL [195], the Akaikie information criteria, AIC [191], and cross-validated log-likelihood, CVLL [193]. The aim of this chapter is to determine which of these statistical model selection criteria can determine the correct number of metastable states from simulations of a model system. The main benefit of the these criteria (except CVLL) are that they require little additional calculation after estimating a model, in contrast to the full Bayes-factor method of reference [188]. This chapter lays the ground-work for application to determining the optimal coarse-grained description of the dynamics of AADH in chapter 6.

1.9 Aromatic amine dehydrogenase

Aromatic amine dehydrogenase (AADH) oxidizes primary aromatic amines, such as tryptamine, into the corresponding aldehyde and ammonia. The rate-limiting step is the proton transfer from a covalently bound Schiff base intermediate to an acceptor aspartate oxygen atom [197]. AADH is notable because it exhibits a large primary kinetic isotope effect: substituting deuterium for the hydrogen being cleaved in tryptamine causes the rate to drop by up to a factor of 55 [197]. A drop in the rate is expected when considering the cleavage of the heavier deuterium atoms as the C—H bond has a larger zero-point energy than the C—D bond, effectively decreasing the height of potential barrier the reaction has to cross. However, if the zero-point energy were the only difference contributing to the difference in rates, a KIE of approximately 8 would be expected [198]. The fact that the observed KIE is almost 7 times as large implies significant quantum mechanical tunneling [197, 199, 200], i.e., at C—H bond distances below the top of the potential energy barrier for the reaction, the proton can transfer to the product state without the need for the kinetic energy required by classical mechanics. In addition to the presence of tunneling indicated by the inflated KIE, the KIE of AADH is independent of temperature, despite the fact that the underlying reaction is dependent on temperature [197].

The motivation for studying AADH and other enzymes such as monoamine dehydrogenase (MADH) [201, 202], soyabeen lipoxygenase (SLO) [203, 204] and DHFR [205, 206] (as well as many others [207]) is that the KIEs are temperature independent and often large in absolute value, which cannot be explained [208] by the dominant explanation of thermally activated reaction rates, namely transition state theory (TST) [209]. These observations have prompted a range of explanations and models going beyond TST [198, 199, 210–212]. Importantly for this thesis these models link the

conformational dynamics of the to observable properties (reaction rates and KIEs) of the enzyme. *Understanding the conformational dynamics of AADH will be important evidence in future debates over the validity of these models.*

In the TST picture the enzyme-substrate complex (for AADH this is the Schiff-base after reaction with tryptamine) undergoes thermal fluctuation from the reactant state, along the reaction coordinate through a transition state, and on to the product state (the oxidized Shiff-base in AADH). The rate at which this happens is proportional to $\exp(-\Delta G^{\mathrm{TS}}/RT)$, where ΔG^{TS} is the free energy difference between the reactant and transition states (the activation free energy), R is the gas constant and T is the temperature. The zero-point energy difference in isotopes changes the value ΔG^{TS}. Tunnelling occurs when the thermal fluctuation along the reaction coordinate brings hydrogen atom close enough to the acceptor atom so that wave-functions of the reactant and product state overlap, effectively lowering the value of ΔG^{TS} [207]. However, as Klinman and Kohen point out [208], this model predicts both rates and KIEs which are temperature dependent, while for AADH the rate determining step is temperature *dependent* while the KIE is temperature *independent*.

The main alternative to TST used to explain enzymatic reactions involving tunneling are 'Marcus-like' (which take their name from their similarity to the Marcus theory of electron transfer [213]) or full-tunneling models, which were originally adapted for hydrogen transfer reactions by Kuznetsov and Ulstrop [214]. These models decouple tunneling from other processes by factorizing the rate into two terms [198, 204, 214] (for an extensive review of the different types of models and their applications see reference [207] and [215, chapters 4, 5 and 6]). The first term describes the process of rearranging the heavy atoms into an state ready for tunneling and is an activated process - i.e., determined by an activation energy. The second term describes the tunneling process and is therefore largely determined by the properties of the atom being transferred (the hydrogen or deuterium atom or ion). This second term describes the probability of of tunneling occurring in terms of fluctuations in the donor-acceptor distance (DAD - the distance moved by the hydrogen atom in the course of the reaction). This may or may not depend on temperature, depending on whether fluctuations in the DAD are necessary for tunneling to occur. This separation allows the model to accommodate both temperature dependent rates of reaction and KIEs which are either temperature dependent or independent [208]. Full tunneling models been applied to experimental results of AADH, MADH and others [207, 208, 210, 216–219] which explain the temperature dependent rates as being largely due to the rearrangement of the enzyme prior to the tunneling process. Once in the necessary configuration, tunneling occurs without the need for further thermal fluctuations, meaning the tunneling rate does not depend on temperature and only on the mass of the transferring hydrogen isotope, giving rise to temperature independent KIEs.

Full-tunneling models are not the only explanation for reaction rates and temperature independent KIEs. In reference [220, 221] the authors argue that by extending transition state theory to include the effect of conformational dynamics, the temperature dependencies of KIEs can be explained without the need for full-tunneling models. Their model posits two conformational sub-states, rapidly

interconverting (relative to the reaction timescale) which both react via different pathways and therefore with different activation energies and different degrees of tunneling. They showed that fitting this model to kinetic data from AADH, MADH, SLO and DHFR, reproduced the temperature dependence of the KIE in all four enzymes, however the fitted parameters have been criticized as being unrealistic [208].

Other models of enzymatic reaction rates have been put forward which incorporate non-equilibrium dynamic motions such as networks of promoting vibrations which couple to the reaction coordinate across the enzyme (see reference [222] for a supportive review of these proposals). In fact, the role of DAD fluctuations in the rate of tunneling driven reactions have prompted some [223, 224] to assume that this implies non-equilibrium effects are needed to explain both the rates of reactions and the catalytic rate enhancement of proteins. However this has been refuted [208] on the grounds that these fluctuations are thermally activated (this will be more thorough described in chapter 6). This critique is part of a larger controversy [207, 208, 220, 221, 225] surrounding proposals for models incorporating non-equilibrium effects in enzyme catalysis. One particular proposal is that certain fast conformational transitions provide the inertia needed for the enzyme to take the reaction to completion [226–228]. While this has been criticized in light of evidence from simulations [229], it provides yet another example of the need to consider the role of conformational dynamics in explaining enzymatic reaction rates.

It is clear from the preceding discussion that the conformational dynamics of an enzyme such as AADH, with its large and temperature independent KIE, will be an important contribution to the debate over the validity of the number of different models of enzyme reactivity. Despite being well studied in other areas, no simulation study has so-far described the conformational dynamics of AADH in the reactant state of its rate determining step.

The aims of **Chapter 6** are thus two-fold. First, this chapter describes molecular dynamics simulations of AADH in its reactant state and uses this data to create a Markov state model description of its conformational dynamics. To do this an set of MSM hyperparameters is optimised and understood in terms of the response surface, utilising the work of chapter 4, in addition a set of sensitivity tests are proposed. The model selection criteria of chapter 5 are used to select the appropriate number of hidden states for a coarse-grained description using HMMs. The second aim is to critically assess the MSM optimisation and model selection criteria using the AADH system as a 'real-world' test. **Chapter 7**, discusses the conclusions of this thesis and sets out concrete steps for further work in this area.

CHAPTER 2

THEORY

Contributions

This chapter represents a summary of existing work and contains no original contributions by the author of this thesis.

2.1 Introduction

This chapter sets out the theory of Markov state models (MSMs) and hidden Markov models (HMMs) to describe the dynamics of biomolecular systems. Table 2.1 summarises the nomenclature used in this chapter.

2.2 Markov processes

Markov state models are now used routinely to quantitatively describe the conformational kinetics and thermodynamics of biomolecular systems using data collected from molecular dynamics simulations [80, 230]. A general, molecular system can be described by a vector of phase space coordinates (momentum and position) as a function of time, $\mathbf{x}(t)$. A thermodynamic ensemble of such systems can be described by a probability density over $\mathbf{x}(t)$, $p(\mathbf{x};t)$ [64]. Modelling a system in thermodynamic equilibrium as a Markov process imposes a number of assumptions on $p(\mathbf{x};t)$ [64]:

1. that there exists a period of time, τ, over which the evolution of the system from a point $\mathbf{x}(t)$ to a new point $\mathbf{y}(t+\tau)$ is dependent only on $\mathbf{x}(t)$, i.e. the joint probability density $p(\mathbf{x},\mathbf{y};\tau)$ is conditional *only* on $\mathbf{x}$:

$$p(\mathbf{x},\mathbf{y};\tau) = \mathbb{P}\left[\mathbf{x}(t+\tau) \in \mathbf{y} + d\mathbf{y} | \mathbf{x}(t) = \mathbf{x}\right]. \tag{2.1}$$

Symbol	Definition
N_A	Number of atoms.
N_T	Number of trajectory frames.
N_C	Number of important continuous features
t	Time index or variable.
$\mathbf{x}/\mathbf{y}(t)$	Point in phase space as function of time.
τ	Markov lag-time.
$p(\mathbf{x},\mathbf{y};\tau)$	Probability of observing $\mathbf{x}$ and then $\mathbf{y}$ a time τ later.
$p(\mathbf{x};t)$	Probability distribution over phase space at time t.
$\mu(\mathbf{x})$	Stationary distribution of system in thermodynamic equilibrium.
$\mathscr{T}(\tau)$	Transfer operator, equation 2.4.
$q(\mathbf{x};t)$	Normalized descriptor of a thermodynamic ensemble. $q(\mathbf{x}) = \frac{p(\mathbf{x})}{\mu(\mathbf{x})}$.
$(\psi_i(\mathbf{x}), \lambda_i)$	Eigenfunctions and eigenvalues of the transfer operator.
r	The number of dominant eigenvalues: $\lambda_{2,\ldots,r} \simeq 1$ and $\frac{\lambda_r}{\lambda_{r+1}} \gg 1$.
n	Number of discrete states/microstates.
$s^i \quad i = 1,\ldots,n$	Indicator functions used to discretize phase space into n states.
$\mathbf{s} = \{s_1,\ldots,s_{N_T}\}$	Trajectory in the basis defined by s^i.
$\mathbf{T}$	Transition matrix. The discrete analogue of the transfer operator.
$\boldsymbol{\pi}$	Stationary distribution of the MSM.
$\mathbf{p}(t)$	State probability vector. Discrete analogue of $p(\mathbf{x};t)$.
$\mathbf{q}(t)$	Normalized state vector. Discrete analogue of $q(\mathbf{x};t)$.
$\mathbf{v}$	Right eigenvectors of $\mathbf{T}$. Discrete analogue of $\psi_i(\mathbf{x})$.
$\mathbf{u}$	Left eigenvectors of $\mathbf{T}$. $\mathbf{u}$ and $\mathbf{v}$ are related by: $u_i = v_i \cdot \pi_i$
$\mathbf{X}$	Data matrix. Coordinates snapshots of trajectory: $\mathbf{X} \in \mathbb{R}^{N_\mathrm{T} \times N_A}$.
$\boldsymbol{\chi}$	Feature matrix. Transformation of $\mathbf{X}$ into important features of system. $\boldsymbol{\chi} \in \mathbb{R}^{N_\mathrm{T} \times N_\mathrm{C}}$.
$\boldsymbol{\chi}'$	TICA transformed feature matrix.
$\mathbf{C}$	Time lagged correlation matrix between states. $C_{ij} = \mathrm{cor}(i,j;\tau)$
$\mathbf{c}$	Count matrix for discrete states. $c_{ij} \propto C_{ij}$.
$\mathbf{S}$	Overlap matrix between either continuous or discrete states. In the discrete basis, $\mathbf{S} = \mathrm{diag}\{\boldsymbol{\pi}\}$.
g	Number of hidden states of a HMM.
$\mathbf{h} = \{h_t\}$	Trajectory of hidden states.
$\widetilde{\mathbf{T}}$	Hidden state transition matrix of a HMM.
$\widetilde{\boldsymbol{\pi}}$	Hidden state stationary distribution of a HMM.
$\mathbf{M}$	Membership matrix of HMM. $M_{ji} = \mathbb{P}(h_t = i \vert s_t = j)$.
$\mathbf{E}$	Emission matrix of HMM. $E_{i,j} = \mathbb{P}(s_t = j \vert h_t = i)$.

TABLE 2.1: IMPORTANT SYMBOLS USED THROUGHOUT THIS CHAPTER.

2. That there are no regions of phase space disconnected from one another, i.e. that the system is *ergodic*. In this case there is a unique stationary distribution, $\mu(\mathbf{x})$. At constant temperature $\mu(\mathbf{x})$ is the Boltzmann distribution.
3. The system is *reversible* and so obeys detailed balance:

$$\mu(\mathbf{x})\,p(\mathbf{x},\mathbf{y};\tau) = \mu(\mathbf{y})\,p(\mathbf{y},\mathbf{x};\tau), \tag{2.2}$$

in other words, the absolute probability of observing a transition from $\mathbf{x}$ to $\mathbf{y}$ (also known as the flux, $F(\mathbf{x},\mathbf{y})$) is the same as that from $\mathbf{y}$ to $\mathbf{x}$.

The dynamics of a Markov process in continuous space is described by the *transfer operator*, $\mathscr{T}(\tau)$, which propagates $q(\mathbf{x};t)$:

$$q(\mathbf{x};t) = \frac{p(\mathbf{x};t)}{\mu(\mathbf{x})}, \tag{2.3}$$

forward in time by [64]:

$$\begin{aligned} q(\mathbf{y};t+\tau) &= \mathscr{T}(\tau)\cdot q(\mathbf{y};t) \\ &= \frac{1}{\mu(\mathbf{y})}\int d\mathbf{x}\,p(\mathbf{x},\mathbf{y};\tau)\mu(\mathbf{x})\,q(\mathbf{x};t). \end{aligned} \tag{2.4}$$

All the kinetic and thermodynamic information of the system is contained within $\mathscr{T}$, its eigenfunctions, $\psi_i(\mathbf{x})$, and eigenvalues λ_i, which, for reversible dynamics, all lie within the interval $-1 < \lambda_i \leq 1$ [85]. The first eigenvector, with $\lambda_1 = 1$, is given by $\psi_1(\mathbf{x}) = \mathbf{1}$ which corresponds to the stationary distribution $\mu(\mathbf{x})$ by virtue of the definition of $q(\mathbf{x})$, equation 2.3 [64]. The remaining eigenvector/eigenvalue pairs shall be assumed to be ordered in decreasing value of λ.

The remaining eigenfunctions, $\psi_{2,3,4\ldots}$, correspond to the relaxation processes which take the system from any initial distribution, $q(\mathbf{x};t=0)$ towards the stationary distribution on a timescale related to its corresponding eigenvalue [64]. This can be seen by writing the time evolution of $q(\mathbf{x};t)$ as:

$$q(\mathbf{x};t+k\tau) = \mathbf{1} + \sum_{i=2}^{\infty} e^{-k\tau/t_i}\left\langle q(\mathbf{x};t),\psi_i(\mathbf{x})\right\rangle_\mu \psi_i(\mathbf{x}), \tag{2.5}$$

where $k = 1,2,3\ldots$ is a time index, and

$$t_i = -\frac{\tau}{\ln|\lambda_i|}, \tag{2.6}$$

are the *implied timescales* for the relaxation process described by $\psi_i(\mathbf{x})$ [64]. The bracketed quantity is the overlap between $q(\mathbf{x})$ and the eigenfunctions:

$$\left\langle q(\mathbf{x};t),\psi_i(\mathbf{x})\right\rangle_\mu = \int d\mathbf{x}\mu(\mathbf{x})\,q(\mathbf{x};t)\psi_i(\mathbf{x}). \tag{2.7}$$

Each term in equation 2.5 will decay exponentially, and in the long time limit, as $k\to\infty$, leave just the first eigenvector, $\psi_1(\mathbf{x}) = \mathbf{1}$, the stationary distribution. If the first r eigenvalues are $\lessapprox 1$ and

are separated from the remaining values by a gap such that $\lambda_r \gg \lambda_{r+1}$, then its possible to truncate equation 2.5, without serious loss of accuracy, to:

$$q(\mathbf{x}, t+k\tau) \simeq \mathbf{1} + \sum_{i=2}^{r} e^{-k\tau/t_i} \left\langle q(\mathbf{x};t), \psi_i(\mathbf{x}) \right\rangle_\mu \psi_i(\mathbf{x}). \tag{2.8}$$

These r eigenfunctions are known as the *dominant* eigenfunctions and they correspond to the slow relaxation processes of the system [64]. The truncation amounts to describing just the slow kinetic processes of the system while ignoring the fast processes. This separation of timescales implies the existence of regions of phase space, partitioned by the dominant eigenfunctions, known as *metastable* states [64].

2.3 Markov state models

Markov state models (MSMs) are discrete models of Markovian dynamics described in previous section [64]. The continuous quantities described above all have discrete analogues which will be described in detail in this section [64]:

- The system is described by a set of n discrete states denoted, $i = 1, \ldots, n$. Instead of the continuous vector $\mathbf{x}(t)$, each trajectory is denoted by a vector of integers $\mathbf{s}$, where each component, s_t, is the state at time t. t is now a discrete quantity, an integer multiple of the time-step, Δt, used to record the coordinates in an MD trajectory: $t = k\Delta t,\ k = 1, 2, \ldots$.
- The system can be described by a probability mass vector, $\mathbf{p}(t)$, instead of a probability density function $p(\mathbf{x};t)$. The ith component of $\mathbf{p}(t)$ is the probability of the system being in state s^i at time t.
- The stationary distribution, $\boldsymbol{\pi}$, is defined by integrating the stationary distribution, $\mu(\mathbf{x})$, over the domain of each discrete state, s^i:

$$\pi_i = \int_{\mathbf{x} \in s^i} \mathrm{d}\mathbf{x} \mu(\mathbf{x})$$

- By analogy with equation 2.3, the system can also be described by $\mathbf{q}(t)$ where $q_i = p_i / \pi_i$.
- The time evolution of $\mathbf{q}(t)$ and $\mathbf{p}(t)$ is determined by the *transition matrix*, $\mathbf{T}(\tau)$:

$$\mathbf{q}(t+\tau) = \mathbf{T}(\tau) \cdot \mathbf{q}(t)$$
$$\mathbf{p}^\top(t+\tau) = \mathbf{p}^\top(t) \cdot \mathbf{T}(\tau)$$

- The eigenfunctions, $\psi(\mathbf{x})$, are now the right eigenvectors of $\mathbf{T}(\tau)$, $\mathbf{v}$, with the same interpretation. The left eigenvectors, $\mathbf{u}$, are related to the right eigenvectors by: $v_i = u_i / \pi_i$.

Creating an MSM starts with the collection of molecular dynamics (MD) data in the form of a set of short trajectories, with configurations saved every Δt seconds. This thesis will consider only canonical ensemble simulations, using both over-damped Langevin dynamics [231, 232] and velocity

re-scaling [233] to maintain temperature. As a consequence, the momentum coordinates will be ignored as a feature for constructing MMs.

If each trajectory has N_T frames and N_A atoms then a trajectory can be represented by a data matrix, $\mathbf{X} \in \mathbb{R}^{N_\mathrm{T} \times 3N_\mathrm{A}}$. The trajectories undergo a series of processing steps on the way to creating an MSM, these are [80, 130]:

1. **Create features:** A set of continuous features, the "essential degrees of freedom" [85], $\chi_i, i \in \{1, \ldots, N_\mathrm{C}\}$ are chosen to capture the slow dynamics of the system:

$$\mathbf{X} \rightarrow \boldsymbol{\chi}, \quad \boldsymbol{\chi} \in \mathbb{R}^{N_T \times N_C}$$

   Examples of continuous features include dihedral angles [234–237], residue contact distances [238–240], root mean square deviation (RMSD) [241, 242], or secondary structure (SS) metrics [243]. Neither SS or RMSD are amenable to further dimensionality reduction (step 2), but can be clustered (step 3) and so are included here.
2. **Dimensionality reduction:** The number of features is reduced still further by transforming $\boldsymbol{\chi}$ into a small number, m, of collective variables.

$$\boldsymbol{\chi} \rightarrow \boldsymbol{\chi}', \quad \boldsymbol{\chi}' \in \mathbb{R}^{N_\mathrm{T} \times m}$$

   This thesis will exclusively use time-lagged independent component analysis (TICA) as a method of dimensionality reduction [81] [89].
3. **Discretization:** Each of the MD frames is assigned to one of n different discrete states using a clustering algorithm such as k-means [244] clustering:

$$\boldsymbol{\chi}' \rightarrow \mathbf{s}, \quad \mathbf{s} \in \mathbb{Z}^{N_\mathrm{T}}$$

4. **MSM estimation**: The transition matrix, $\mathbf{T}$, is estimated by counting transitions between discrete states separated by a time τ.
5. **Coarse-graining:** The MSM is then coarse-grained by grouping the n microstates into g macrostates states. This thesis will exclusively consider hidden Markov models (HMMs) as a method for doing this.

2.3.1 Create features

The choice of continuous feature χ may be determined or at least strongly suggested by the question being asked and/or from prior knowledge of the system. However, since the introduction of variational scoring rules such as the GMRQ [94] and VAMP [97], they are typically chosen using cross-validation. This is discussed in depth in chapter 4.

2.3.2 Dimensionality reduction with TICA

Dimensionality reduction using TICA was introduced as a preprocessing step for improving MSM construction [81, 89]. The result of a time-lagged independent component analysis is an estimate of the eigenvectors of the transfer operator. So if MD trajectories have been projected onto the contact distances of the protein, then the TICA eigenvectors are linear combinations of these distances such that the first component represents the equilibrium distribution, the second component represents the slowest relaxation process and so on. These are approximations to the true transfer operator eigenvectors because the true eigenvectors, in general, are non-linear in such features. If TICA is performed then the feature trajectories can be projected onto the first m TICA components (where m must be chosen), creating a new, smaller set of features. There are two advantages to this. First, the fact that the number of features has been reduced to m dimensions, means that clustering will be more efficient. This is because the amount of computation required by k-means clustering, a popular approach to clustering, scales with the number of dimension [87], as was discussed in section 1.3 of the introduction to this thesis. The second, more important advantage, is that if the TICA components are scaled by their eigenvalues (a process called kinetic mapping) then distances in the space of these components approximate *kinetic distances* [90]. The kinetic distance between two conformations correlates with how slowly they interconvert. This means that geometric similarity now implies (at least approximately) kinetic similarity. Geometrically discretizing this space means that molecular dynamics frames which are all mapped to the same discrete state, will have similar rates of interconversion to other discrete states. This fulfills the main assumption behind approximating the dynamics of a system with a discrete master equation (see equation 1.1 of the introduction) [62, 63].

The TICA operator is a result of variationally optimising a set of continuous basis functions, χ_i, to estimate the eigenfunctions, ψ_i, of the transfer operator. The method can be summarised as follows: a trial function, f, is expanded in the basis χ_i:

$$f(\mathbf{x}) = \sum_{i}^{N_C} a_i \chi_i(\mathbf{x}) \tag{2.9}$$

The variational principle for operators with bounded eigenvalues (such as the transfer operator) states that any approximate eigenfunction will have eigenvalues less than the true eigenvalues [245]. So choosing the coefficients a_i to maximize the eigenvalues from using equation 2.9 will be the closest approximation to the true eigenfunctions that can be achieved with linear combinations of basis functions. Using the method of Lagrange multipliers to maximize the eigenvalues using $f(\mathbf{x})$ results in the following generalized eigenvalue equation [91]:

$$\mathbf{Ca} = \lambda \mathbf{Sa} \tag{2.10}$$

Solving this equation requires estimating the matrix elements of $\mathbf{C}$ and $\mathbf{S}$. The elements of $\mathbf{C}$ are the

time-lagged correlation functions (cor) between χ_i and χ_j [91]:

$$\begin{aligned} C_{ij} &= \mathrm{cor}\left(\chi_i, \chi_j, \tau\right) \\ &= \int\int \chi_i(\mathbf{z}) \mathbb{P}\left(\mathbf{x}(t+\tau) = \mathbf{z} \,|\, \mathbf{x}(t) = \mathbf{y}\right) \\ &\quad \times \chi_j(\mathbf{y}) \mathbb{P}\left(\mathbf{x}(t) = \mathbf{y}\right) \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{z}, \end{aligned} \tag{2.11}$$

which can be estimated from the MD trajectories. The matrix $\mathbf{S}$ is the overlap matrix whose elements are the scalar product between the basis functions, weighted by the stationary distribution [91]:

$$S_{ij} = \int \chi_i(\mathbf{x}) \chi_j(\mathbf{x}) \mu(\mathbf{x})^{-1} \mathrm{d}\mathbf{x} \tag{2.12}$$

Having solved for the TICA eigenvectors, keep the first m columns of $\mathbf{a}$ and use this to transform feature matrix $\boldsymbol{\chi}$ [81, 89]:

$$\boldsymbol{\chi}' = \boldsymbol{\chi} \cdot [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_m] \tag{2.13}$$

2.3.3 Discretization

Discretization is performed on the m dimensional feature matrix $\boldsymbol{\chi}'$ using a clustering algorithm such as k-means or Ward clustering [246]. Each region of feature space, S_i, is assigned uniquely to the microstate i through the associated indicator function, $s^i\left(\boldsymbol{\chi}'\right)$ [64]:

$$s^i\left(\boldsymbol{\chi}'\right) = \begin{cases} 1, & \boldsymbol{\chi}' \in S_i \\ 0, & \boldsymbol{\chi}' \notin S_i \end{cases} \tag{2.14}$$

To avoid introducing too many symbols, $s^i\left(\boldsymbol{\chi}'\right)$ denotes the indicator functions, while $\mathbf{s}$ is the MD trajectory in the indicator function basis. The individual components of $\mathbf{s}$ will be denoted s_t where t is a time index. To highlight the time series nature of $\mathbf{s}$ it will sometimes denoted $\{s_1, s_2, \ldots\}$.

2.3.4 MSM estimation

MSM estimation is analogous to TICA estimation but with indicator basis functions s^i replacing the continuous basis functions χ_i in equation 2.9 [91]. The result of the optimisation of coefficients a_i results in the same generalized eigenvalue expression, equation 2.10. Solving this equation is simplified by the fact that the indicator functions are orthogonal and so $S_{ij} = 0$ if $i \neq j$ and $S_{ii} = \pi_i$. Equation 2.10 then becomes [64, 91]:

$$\mathbf{C}\mathbf{a} = \lambda \boldsymbol{\Pi} \mathbf{a} \tag{2.15}$$

$$\mathbf{T}\mathbf{a} = \lambda \mathbf{a} \tag{2.16}$$

where $\boldsymbol{\Pi} = \mathrm{diag}\{\pi_1, \ldots, \pi_n\}$ and $\mathbf{T} = \boldsymbol{\Pi}^{-1}\mathbf{C}$. In other words:

$$T_{ij}(\tau) = \frac{\mathrm{cor}(s_i, s_j, \tau)}{\pi_i} \tag{2.17}$$

For systems in thermodynamic equilibrium, the transition matrix elements must be estimated such that they respect the assumptions of i) reversibility and detailed balance, and ii) that the ergodicity i.e. that each state is connected to all other states [64, 91]. Two methods exist for estimating T_{ij}: maximum likelihood estimation, MLE [64, 247], and Bayesian estimation [248]. This work will make use of both: MLE will be used for model selection while Bayesian optimisation will be used when error estimates are required. Detailed explanations of both techniques can be in reference [248] but certain concepts need highlighting here: i) maximum likelihood estimation, ii) count matrices and ergodicity, iii) Bayesian estimation.

2.3.4.1 Maximum likelihood estimation

Parameter estimation through maximum likelihood estimation proceeds by first modelling the probability of observing the data $\{s_1, s_2, s_3, \ldots\}$ given *fixed parameters*, T_{ij} [249]. For a MSM with $\tau = 1$ this is [248]:

$$\mathbb{P}(\mathbf{s}|\mathbf{T}) \propto \prod_{t=1}^{N_\mathrm{T}} T_{s_t, s_{t+1}} \tag{2.18}$$

The likelihood, $\mathscr{L}(\mathbf{T}|\mathbf{s})$, is equal $\mathbb{P}(\mathbf{s}|\mathbf{T})$ but treats the the *data as fixed* and the *parameters as varying*[1] [249]. Equation 2.18 can be re-written in terms of the count matrix, c_{ij}, which counts all the observed transitions between states i and j [248]:

$$\mathscr{L}(\mathbf{T}|\mathbf{s}) = \mathbb{P}(\mathbf{s}|\mathbf{T}) \tag{2.19}$$

$$\propto \prod_{t=1}^{n_f} T_{s_t, s_{t+1}} \tag{2.20}$$

$$\propto \prod_i \Pi_j T_{i,j}^{c_{ij}} \tag{2.21}$$

$$= \mathbb{P}(\mathbf{c}|\mathbf{T}) \tag{2.22}$$

So to calculate the maximum likelihood estimates (MLE) for T_{ij}, the count matrix $\mathbf{c}$ is estimated from the trajectories and then T_{ij} are varied until a maximum value of $\mathscr{L}$ is found [249]. For reversible MSMs an iterative scheme is used which ensures $\mathbf{T}$ satisfies detailed balance [248].

2.3.4.2 Count matrix

When the Markov lag-time is greater than the time-lag between frames, i.e., $\tau > 1 \cdot \Delta t$, then there is ambiguity as to how the count matrix should be calculated. For accurate confidence intervals for the parameters T_{ij} each pair of states, $(s_t, s_{t+\tau})$ should be independent [248]. The two extreme methods for counting are *sliding window* and *sample count* [248].

Sliding window counts all pairs of frames separated by τ:

$$(s_0, s_\tau),\ (s_1, s_{1+\tau}),\ (s_2, s_{2+\tau})$$

[1] The integral of the likelihood over the parameter space can be greater 1 hence it is not a probability.

For example, a trajectory with $N_{\mathrm{T}} = 100$ frames, with $\tau = 2$ has $\sum_{ij} c_{ij} = 98$ observed transitions. If, as is usually the case with MD data, the observations are correlated at $t < \tau$, the confidence intervals generated will be too narrow, i.e., optimistic.

The *sample count* method uses the only the states separated by τ:

$$(s_0, s_\tau),\ (s_\tau, s_{2\tau}),\ (s_{2\tau}, s_{3\tau})$$

Continuing the previous example, $\sum_{ij} c_{ij} = 49$, a factor of 2 different to the sliding window method. This leads to the confidence intervals being over-estimated. A third method, *effective counting*, uses the sliding window method but scales count matrix elements by a factor, I_{ij}, to account for correlation between the observations [248, 250]. Continuing the previous example, if the observations separated by $t < \tau$ are weakly correlated then the effective count matrix will be close to the sliding window estimate and $I_{ij} \lesssim 1$; if the observations are highly correlated the count matrix will be closer to the sample count method and $I_{ij} < 1$. This method has been shown to give more accurate confidence intervals [248].

An MSM can only describe transitions between groups of microstates that are ergodic. For example, with 10 microstates, if no transitions are observed between a group consisting of microstates 1 and 2 and a group consisting of microstates 3–10 (but transitions are observed between all other pairs), then an MSM describes either the first group (a 2-state MSM results), or the second group (a 8-state MSM results). However, the method of counting transitions between states, will determine whether or not those states are ergodic. There is a path between states a and z if there is a set of intermediate states, $b, c, \ldots, y$ such that $c_{a,b} > 0,\ c_{b,c} > 0,\ \ldots,\ c_{y,z} > 0$. Two states are strongly or reversibly connected if there is a path from $a \to z$ and from $z \to a$. The full set of states are ergodic if each state pair of states is strongly connected [251]. The sliding window procedure utilises more transitions than the sample count method, and will generally increase the size of the strongly connected set. In all estimation procedures used in this thesis, if the full set of state are not strongly connected, then a subset of states are used such that all states within the subset are strongly connected.

2.3.4.3 Bayesian estimation

Bayesian estimation uses not just the likelihood, $\mathbb{P}(\mathbf{c}|\mathbf{T})$, but the prior probability of the parameters, $\mathbb{P}(\mathbf{T})$, to estimate the posterior probability of the parameters given the data, $\mathbb{P}(\mathbf{T}|\mathbf{c})$, using Bayes' rule [252]:

$$\mathbb{P}(\mathbf{T}|\mathbf{c}) \propto \mathbb{P}(\mathbf{c}|\mathbf{T}) \cdot \mathbb{P}(\mathbf{T}) \tag{2.23}$$

Rather than point estimates of the parameters, samples are drawn from the whole posterior distribution for each parameter, T_{ij}, using Markov chain Monte Carlo [252]. These samples can then be used to estimate the distribution, and hence error, of any quantity determined by the transition matrix, e.g., eigenvalues or implied timescales [252]. The details of the sampling procedures used for

estimation of MSM in this work can be found in reference [248] and are implemented in the Python package PyEMMA (version 2.5) [253], however two points should be highlighted here. First, the count matrices used are estimated using the effective count method. Second, the prior distribution for the transition matrix elements are defined over a reversible matrix $\mathbf{X}$, where $X_{ij} \propto T_{ij}\pi_i$, and is given by [248]:

$$\mathbb{P}(\mathbf{X}) \propto \Pi_{i \geq j} x_{ij}^{-1}. \tag{2.24}$$

This is to ensure the prior is consistent with detailed balance. As a result the posterior transition matrix elements have the same connectivity structure as the count matrix. i.e., if $c_{ij} = 0$ then for any posterior sample $T_{ij} = 0$ [248].

A method for testing the convergence of the MCMC sampling procedure is to calculate the rank-normalized $\hat{R}$ statistic using four independent sampling chains [254]:

$$\hat{R} = \frac{\hat{V}}{W}. \tag{2.25}$$

Here $\hat{V}$ is the pooled variance of the four chains after rank normalization (i.e., sampled values are ranked and then the rank converted to a standard normal distributed variable) and W is the within-chain variance. The recommended threshold value for convergence in reference [254] is 1.01, i.e if the sampling is converged then $\widehat{R} < 1.01$. The more stringent convergence tests in reference [254] are not considered here due to the exploratory nature of this thesis.

2.3.5 Choice of Markov lag-time

The lag time, τ, is not a hyperparameter of the model but rather a part of the model specification; it defines the temporal resolution of the description of the dynamics. For large values of τ the truncation in equation 2.8 becomes more accurate, or, put another way, the Markov assumption becomes more plausible [64]. In addition, for small values of τ the correlation between the observations will mean that the transition matrix is non-Markovian, i.e., does not obey the Chapman-Kolmogorov equation [249]:

$$[\mathbf{T}(\tau)]^k = \mathbf{T}(k\tau) \tag{2.26}$$

However, if τ is too large then there may be processes of interest with timescales, $t_i < \tau$ which will not be captured [64].

As a compromise, τ is chosen by inspection of the data through an *implied timescale plot* [64, 255]. Starting from a given discrete trajectory, $\{s_t\}$ the implied timescales, t_i, are estimated for a range of different values of τ. The lag time is chosen as the smallest τ such that $\frac{dt_i}{d\tau} \simeq 0$, where the range of i is determined by the number of number slow relaxation processes.

2.3.6 Choice of hyperparameters

The hyperparameters of the MSM are the modelling choices which transform the raw MD data into the n discrete states, s^i, used in MSM estimation. This work will consider only the following hyperparameters:

1. χ the continuous feature,
2. τ' the TICA lag time (τ' can be different from τ),
3. m the number of retained TICA components,
4. n the number of cluster centers.

In principle, there are a large number of modelling choices which will have an effect on the final definition of the discrete states. These will described in the relevant sections later in this thesis. However, two important choices which will be kept fixed are:

1. the scaling of the TICA components. Each component will be scaled by their eigenvalues (the "kinetic mapping" option in PyEMMA [253]);
2. the k-means [244] algorithm will be used to cluster the TICA transformed trajectories into the discrete states.

Kinetic mapping is recommended for TICA dimensionality reduction [90] because it maps geometric distances to kinetic distances (section 2.3.2). As already discussed in section 1.3 of the introduction to this thesis, a recent study looked at the quality of MSM models, created with different MSM hyperparameters [92], of twelve fast-folding proteins. The metric they used to judge model quality was the generalized matrix Rayleigh coefficient, which is discussed in full below, but essentially measures how accurately the model describes the slow eigenvectors of the true transition matrix. They found that on average using kinetic mapping with TICA increases the quality of the model compared to using TICA without kinetic mapping (which is in turn better than using PCA or no preprocessing). Kinetic mapping mitigated the degradation in model quality introduced by including too many TICA components. However, this effect was smallest when using a small number of TICA components and for seven of the 12 optimum hyperparameters (the optimum for each protein) kinetic mapping was not used. This may be due to the small number of scoring iterations used to estimate the model score: only five iterations were used, later studies on the same systems used 50 [130].

The choice of clustering algorithm has been investigated in reference [246] where the authors looked at the quality of MSMs, created using different clustering algorithms, of three of the twelve fast-folding proteins investigated in reference [92]. They found that k-means and Ward clustering performed equivalently well when judged with the same model score used in reference [92]. K-means will be used in this work because of its good performance and the fact that it is implemented in the MSM package PyEMMA [253]. Fixing both the kinetic mapping and clustering algorithm reflects an arbitrary decision to keep the number of hyperparameters to a minimum.

The variational principle, used previously to optimize the coefficients in the basis function expansion, equation 2.9, can also be used to optimize the basis functions themselves [92, 130]. In order to do this objective functions known as the VAMP (variational approach to Markov processes) scores [97] were derived which reach their upper limit when the resulting eigenfunctions become the exact eigenfunctions of the transfer operator. The VAMP scores generalize previous work on variationally optimising basis functions using the generalized matrix Rayleigh coefficient, GMRQ [94]. The VAMP scores are a family of scores, parameterized by an integer w [97]:

$$\text{VAMP}-w = \sum_{i=1}^{r} (\lambda_i)^w, \tag{2.27}$$

where r is the number of dominant eigenvalues of the transition matrix i.e. the r in equation 2.8. The GMRQ corresponds to $w = 1$ and has been used to optimize parameters for MSMs of protein folding [92]. This work will follow the suggestion in reference [97] and a recent study which looked at selecting appropriate features for protein folding [130]. They used the VAMP score with $w = 2$ (VAMP-2), which is equivalent to maximizing the kinetic variance [97]. Maximizing the VAMP-2 score is then equivalent to choosing a basis set which most accurately describes the the slow dynamic processes of the system (see the earlier discussion of kinetic variance in section 2.3.2).

In order to use the VAMP-2 score the following steps can be used [130]:

1. select a set of hyperparameters to create a MSM basis, s^i,
2. estimate the MSM transition matrix,
3. calculate the VAMP-2 score,
4. repeat steps (1) - (3) a number of times until the VAMP-2 score converges.

However, using equation 2.27, will tend to produce overly complex basis functions that fit to noisy fluctuations in the data, rather than the 'true' relaxation processes [94, 97, 130], a process known as over-fitting [150].

Two popular techniques for mitigating over-fitting [150] are the bootstrap [256] and cross-validation, CV [95]. The bootstrap consists of drawing samples (typically 100s or 1000s) from the data, with replacement, to form new data sets. Models are estimated and scored on these data. The distribution of model scores is used to calculate their mean and confidence intervals. Models can then be compared using these mean scores, which take into account an estimate of the sampling uncertainty. The idea of CV is to estimate the model parameters on a subset of the data, the training data $\mathscr{D}^{\text{train}}$, and then evaluate the performance of the parameters on the remaining data $\mathscr{D}^{\text{test}}$ [150]. K-fold CV is popular for MSMs [94, 130]. In this method, the model is trained on $K-1$ disjoint 'folds', where a fold is $1/K$ of the total data, e.g., if there are 100 MD trajectories, a fold will be $100/K$ trajectories. The score is then evaluated on the single held out fold and the average score over the K folds is used to evaluate the model. However, many different methods exist [95] and this work will use 50:50 shuffle split CV, which has been used previously for MSMs [92, 257]. It is equivalent to repeated 2-Fold CV with a random permutation of the data in between each evaluation and is described in algorithm 1.

Algorithm 1: 50:50 shuffle split cross-validation.

Data: N independent trajectories: $\mathscr{D} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N]$
Input: hyperparameter set: θ
Input: estimator of model parameters: $P = P(\theta; \mathscr{D})$
Input: scoring function: $CV(\theta) = CV(P; \mathscr{D})$
Parameter: number of iterations: J

for $i \leftarrow 1$ **to** J **do**
  permute ("shuffle") location of trajectories in $\mathscr{D}$: e.g., $\mathscr{D} \rightarrow [\mathbf{s}_5, \mathbf{s}_{17}, \ldots, \mathbf{s}_2]$
  split trajectories into two equal size, disjoint sets: $\mathscr{D} \rightarrow \mathscr{D}^{\text{train}}, \quad \mathscr{D}^{\text{test}}$
  estimate model parameters on training data: $P^{\text{train}} = P(\theta, \mathscr{D}^{\text{train}})$
  score model on test data: $CV_i^{\text{test}} = CV_i(P^{\text{train}}; \mathscr{D}^{\text{test}})$

Result: Cross-validated score: $\text{MCV}(\theta) = \frac{1}{J} \sum_i CV^{\text{test}}(\theta)$

It is important to note that for CV to judge the performance of all the hyperparameters, the 'estimator of model parameters' must take as its input the *raw* data [150]. For the case of MSMs, the estimator takes the cartesian coordinates, applies TICA, clusters trajectories, and then estimates the transition matrix elements. The scoring function, CV, for the VAMP-2 score is given by [97]:

$$CV\left(\mathbf{C}, \mathbf{S}, \mathbf{v} \mid \mathscr{D}_{\text{test}}\right) = \left\| \left(\mathbf{v}^\top \mathbf{S}^{\text{test}} \mathbf{v}\right)^{-\frac{1}{2}} \left(\mathbf{v}^\top \mathbf{C}^{\text{test}} \mathbf{v}\right) \left(\mathbf{v}^\top \mathbf{S}^{\text{test}} \mathbf{v}\right)^{-\frac{1}{2}} \right\|_2^2 \tag{2.28}$$

where $\|\mathbf{A}\|_2^2$ is the square of the Frobenius norm of the matrix $\mathbf{A}$ [97]. The matrices $\mathbf{S}^{\text{test}}$ and $\mathbf{C}^{\text{test}}$ are the overlap and time-lagged correlation matrices estimated on the test data, while $\mathbf{v}$ are the eigenvectors of the transition matrix estimated using the training data.

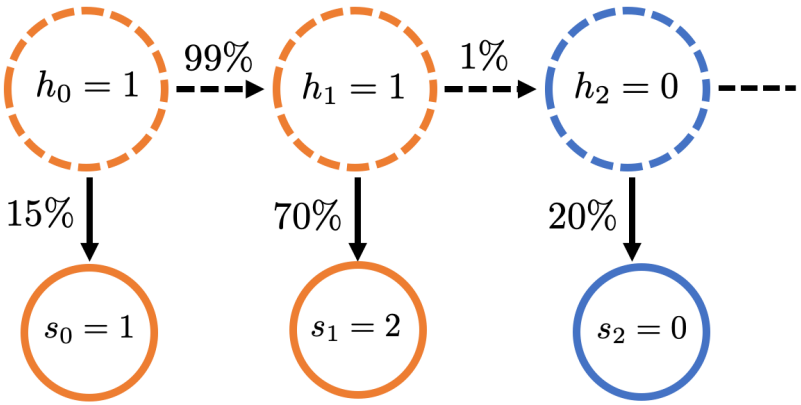
2.4 Coarse-graining with hidden Markov models

2.4.1 HMM definition

Hidden Markov models (HMMs) are models of a dynamic process consisting of the following elements[2] [171]:

1. A number of hidden states, g. These are not observed in the data used to train the model.
2. Hidden-state to hidden-state transition probabilities, $\mathbb{P}(h_{t+1}|h_t)$, where h_t is the hidden state at time t. These are encoded in the hidden state transition matrix, $\widetilde{\mathbf{T}} \in \mathbb{R}^{g \times g}$. This has the same interpretation as the MSM transition matrix, $\mathbf{T}$. In keeping with the notation of reference [169], the $\widetilde{\ }$ pertains hidden quantities.
3. A number of observed states, n. These are the observations used to train the model.
4. Probabilities of seeing the observed states, given a hidden state. These encoded in the 'emission matrix', $\mathbf{E}$, where $E_{ij} = \mathbb{P}(s = j|h = i)$.

[2] This treatment focuses exclusively on discrete hidden Markov processes.



$$\widetilde{\mathbf{T}} = \begin{bmatrix} 91\% & 9\% \\ 1\% & 99\% \end{bmatrix} \qquad M_{ji} = \frac{\widetilde{\pi}_j}{\pi_i} E_{ij}$$

$$\widetilde{\boldsymbol{\pi}} = \begin{bmatrix} 10\% & 90\% \end{bmatrix}$$

$$\mathbf{E} = \begin{bmatrix} 20\% & 60\% & 20\% & 0\% \\ 0\% & 15\% & 70\% & 15\% \end{bmatrix} \qquad \mathbf{M} = \begin{bmatrix} 100\% & 0\% \\ 31\% & 69\% \\ 3\% & 97\% \\ 0\% & 100\% \end{bmatrix}$$

$$\boldsymbol{\pi} = \begin{bmatrix} 2\% & 19.5\% & 19.5\% & 13.5\% \end{bmatrix}$$

FIGURE 2.1: EXAMPLE HIDDEN MARKOV MODEL. A schematic representation of a HMM with $g = 2$ hidden states and $n = 4$ observed states. The dashed circles represent the hidden states, the solid circles the observed states. Transition probabilities (from $\widetilde{\mathbf{T}}$) between hidden states label the dashed arrows; emission probabilities (from $\mathbf{E}$) label the solid arrows.

5. An initial distribution of hidden states, $\widetilde{\boldsymbol{\pi}}'$.

An example HMM is shown in figure 2.1. The hidden states are shown in dashed circles and the hidden-state to hidden-state transition probability label the dashed arrows. These are taken from the transition matrix $\widetilde{\mathbf{T}}$ shown. The stationary distribution, $\widetilde{\boldsymbol{\pi}}$, ensures detailed balance:

$$\widetilde{T}_{1,2} \times \widetilde{\pi}_1 = \widetilde{T}_{2,1} \times \widetilde{\pi}_2$$
$$0.09 \times 0.1 = 0.9 \times 0.01$$

While in each hidden state the system emits to a observed state shown as solid circles and the emission probabilities label the solid arrows. These are taken from the emission matrix $\mathbf{E}$ shown. Other quantities of interest are the observed state distribution $\boldsymbol{\pi}$ which is related to the emission and stationary distribution by $\pi_j = \sum_i E_{ij} \widetilde{\pi}_i$; and the membership matrix, $\mathbf{M}$. The membership matrix encodes the probability of the system being in an hidden state given the observed state [169], i.e., $M_{ji} = \mathbb{P}(h = i | s = j)$. This can determined from the stationary distributions of the hidden states and

observed states, the emission matrix and the laws of probability:

$$\mathbb{P}(h=i|s=j) = \frac{\mathbb{P}(h=i)}{\mathbb{P}(s=j)}\mathbb{P}(s=j|h=i) \tag{2.29}$$

$$M_{ji} = \frac{\widetilde{\pi}_j}{\pi_i}E_{ij} \tag{2.30}$$

The eigenvalues and eigenvectors of the hidden transition matrix have a similar interpretation as the eigenvectors and eigenvalues of an MSM, i.e., they are the relaxation processes and associated timescales, respectively, of the hidden states.

2.4.2 Coarse-graining procedure

HMMs have been proposed as a method for modelling biomolecular dynamics by coarse-graining a MSM [169]. This coarse-graining is accurate under the following assumptions [169]:

1. The underlying dynamics of the system are Markovian, i.e. they can be modelled by a transfer operator, $\mathscr{T}(\tau)$ (equation 2.4).
2. There is a gap between the r'th and $r+1$'th eigenvalues of $\mathscr{T}(\tau)$ i.e. $\frac{\lambda_r}{\lambda_{r+1}} \gg 1$.
3. The r dominant eigenfunctions partition the stationary distribution into r metastable states. The probability of the system being in the boundary between these sets is negligible.

The process for coarse-graining an MSM is as follows (adapted from algorithm 1 in [169]):

1. **Estimate an MSM**. Using the process described in section 2.3, transform the raw MD trajectories, $\{\mathbf{X}_1, \mathbf{X}_2, \ldots\}$, into trajectories of discrete states, $\{\mathbf{s}_1, \mathbf{s}_2, \ldots\}$. Using these discrete trajectories estimate the MSM transition matrix, $\mathbf{T}$ at a lag-time τ.
2. **Determine number of metastable states**. The number of metastable states, r, is determined by looking for a gap in the eigenvalues of $\mathbf{T}$, such that $\frac{\lambda_r}{\lambda_{r+1}} \gg 1$. *The number of hidden states of the HMM, g, is equal to the number of metastable states, r.*
3. **Coarse-grain the transition matrix**. Estimate initial HMM parameters, $\widetilde{\mathbf{T}}_0$, $\mathbf{E}_0$ and $\widetilde{\boldsymbol{\pi}}'_0$, using robust PCCA [159].
4. **Estimate the HMM**. Optimise the HMM parameters (maximize the likelihood) using the Baum-Welch algorithm [170, 258], which is described in the next section.

Robust PCCA (or PCCA+) effectively utilises the sign structure of the dominant eigenvectors of the MSM in the microstate basis to assign microstates to metastable states. For example, if there are $r=3$ dominant eigenvalues (including the $\lambda=1$ eigenvalue corresponding to the stationary distribution) identified in step 2, then PCCA+ will partition the microstates into three metastable states. The method for assigning microstates to metastable states is as follows (this follows the description given in reference [168] assuming grouping into three metastable states): for each microstate i, create a vector $\mathbf{v}^i = (q_1^i, q_2^i, q_3^i)$ where q_k are the values of the kth eigenvector for that microstate. From among all the $\mathbf{v}^i$ choose three that are the most distinct from one another

(using the Gram-Schmidt orthonormalisation algorithm). These three states are assigned to the three metastable states. Assign the remaining microstates to either of these three metastable states based on their similarity to the three representative vectors $\mathbf{v}^i$.

The g eigenvalues and $g-1$ relaxation processes of the HMM will be the coarse-grained equivalent of the eigenvalues and eigenvectors measured in the MSM basis [169].

2.4.3 HMM estimation

The Baum-Welch algorithm as used for coarse-graining reversible MSMs is given in detail in reference [169]. The algorithm is sketched in algorithm 2 to introduce some of the important quantities used later in this thesis. In particular, the 'forward' part of the algorithm calculates the $\alpha_i(t)$ variable which is the probability of arriving at hidden state i at time t *and* seeing the actual observed trajectory up to that point [171]. Summing this value for $t = N_{\mathrm{T}}$ over the hidden states gives the probability of the observed trajectory given the model parameters [171]. This is equal to the likelihood of the parameters given the observed trajectory, $\mathbb{P}(\{s_t\}|\theta) = \mathscr{L}(\theta|\{s_t\})$ [171].

In addition to the maximum likelihood estimation of the parameters, Bayesian estimation can be used. The details of the implementation used in PyEMMA (version 2.5) [253] can be found in reference [259] and the references therein, but the broad outline for estimating a Bayesian HMM with g hidden states is as follows: First, an MSM is estimated and the implied timescales, $t_2, t_3, \ldots$, saved. Second, the trajectories are sub-sampled, or strided, to account for the correlation between observed states. The striding is by a factor Δt given by [253]:

$$\Delta t = \min\left(\tau, 2 \cdot t_{g+1}\right). \tag{2.31}$$

i.e., if $g = 5$ the HMM will capture the first five timescales in the full MSM basis. The 6th implied timescale measured in the MSM basis will the slowest relaxation timescale but which is nevertheless considered too fast to be included in the HMM. This is analogous to the expression given in reference [250]. This means only the following transition are counted:

$$\{(s_0, s_\tau),\ (s_{\Delta t}, s_{\Delta t+\tau})\ \ldots\}. \tag{2.32}$$

Third, a maximum likelihood HMM is estimated and is used to i) determine the largest strongly connected set of hidden states and ii) define the prior distribution of $\widetilde{\boldsymbol{\pi}}'$. The count matrix of the maximum likelihood model is determined by ξ_{ij} from the Baum-Welch algorithm. The largest connected set is defined the same way as for the MSM case.

Fourth, the parameters are sampled using MCMC. The prior function for $\widetilde{\mathbf{T}}$ is given by equation 2.24 and for $\widetilde{\boldsymbol{\pi}}'$ is given by:

$$\widetilde{\boldsymbol{\pi}}' \sim \prod_i \widetilde{\pi}_{0,i}^{a_i + n_i - 1}, \tag{2.33}$$

where a_i are the initial distribution from the maximum likelihood model and n_i is the population of each hidden state at each sampling step in the MCMC algorithm. No priors for the emission

Algorithm 2: The Baum-Welch algorithm.

Data: Initial HMM parameters: $\theta_0 = (\widetilde{\mathbf{T}}_0, \mathbf{E}_0, \widetilde{\boldsymbol{\pi}}'_0)$
Data: observed state trajectory: $\{s_t\}$, $t = 1, \ldots, N_\mathrm{T}$
Parameter: likelihood tolerance: ϵ

begin
  ll_0, θ, continue $\longleftarrow$ 0, θ_0, `True`
  while continue **do**
    Forward procedure
    Calculate the probability of being in hidden state i at time t and seeing the partial trajectory $s_1, \ldots, s_{t'}$, given the model parameters:

$$\alpha_i(t) = \mathbb{P}(\{s_1, \ldots, s_t\} | h_t = i, \theta)$$

    Backward procedure
    Calculate the probability of seeing the partial trajectory $s_{t+1}, \ldots, s_{N_\mathrm{T}}$, given being in hidden state i and the model parameters, θ:

$$\beta_i(t) = \mathbb{P}(\{s_{t+1}, \ldots, s_{N_\mathrm{T}}\} | h_t = i, \theta)$$

    Update model parameters
    Calculate probability of being in hidden state i at time t given the entire trajectory $\{s_t\}$ and θ:

$$\gamma_i(t) = \mathbb{P}(h_t = i | \{s_t\}, \theta) = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{g} \alpha_j(t)\beta_j(t)}$$

    Calculate probability of begin in hidden state i and time t and transitioning to state j at time $t+1$ given the entire trajectory $\{s_t\}$ and θ:

$$\xi_{ij}(t) = \mathbb{P}(h_t = i, h_{t+1} = j | \{s_t\}, \theta) = \frac{\alpha_i(t) T_{ij} \beta_j(t+1) E_{j,s_{t+1}}}{\sum_{k,l}^{g} \alpha_k(t) T_{kl} \beta_l(t+1) E_{l,s_{t+1}}}$$

    Update parameters using:

$$\begin{aligned}
\widetilde{\pi}'_i &\longleftarrow \gamma_i(t=1) \\
\widetilde{T}_{ij} &\longleftarrow \frac{\sum_{t=1}^{N_T-1} \xi_{ij}(t)}{\sum_{t=1}^{N_T-1} \gamma_i(t)} \\
E_{ij} &\longleftarrow \frac{\sum_{t=1}^{N_T} 1_{s_t=j} \gamma_i(t)}{\sum_{t=1}^{T} \gamma_i(t)} \\
\theta &\longleftarrow (\widetilde{\boldsymbol{\pi}}', \widetilde{\mathbf{T}}, \mathbf{E})
\end{aligned}$$

    Calculate log-likelihood
    $\mathrm{ll}' = \log\left(\sum_i^g \alpha_i(N_T)\right)$
    if $\mathrm{ll}' - \mathrm{ll} < \epsilon$ **then**
      continue $\longleftarrow$ `False`
    $\mathrm{ll} \longleftarrow \mathrm{ll}'$

distributions are used, instead the values of are determined from the other sampled quantities. Convergence of the parameters is performed in the same manner as for the MSM, i.e., through sampling independent chains and calculating the $\hat{R}$ statistic.

2.5 Markov model validation

Markov model validation for both MSMs and HMM is performed by checking to see if the Chapman-Kolmogorov equation,

$$[\mathbf{T}(\tau)]^k \approx \mathbf{T}(k\tau), \tag{2.34}$$

holds to within sampling error for a range of k. This check is called the *Chapman-Kolmogorov test* (CK test) [64]. The left-hand side is a transition matrix *predicted at time* $k\tau$ from the matrix estimated at time τ. The right hand-side is the transition matrix *estimated at time* $k\tau$. This equality will hold exactly for a discrete Markov process (i.e., one for which there was no error in creating the discretized states, s^i [64]). For an MSM, testing equality of two $n \times n$ transition matrices requires n^2 comparisons which is both computationally intractable and would result in large uncertainties [64]. Instead, the MSM is coarse-grained as a HMM and the CK test adapted as follows [64]:

1. For each value of $k = 1, 2, 3, \ldots$ calculate $\widetilde{\mathbf{T}}(k\tau)$
2. Using an initial state vector $\mathbf{p}(0)$ and $\widetilde{\mathbf{T}}(\tau)$ predict the values of $\mathbf{p}(k\tau)$ using:

$$\mathbf{p}^\top(k\tau)_{\mathrm{HMM}} = \mathbf{p}(0)^\top [\widetilde{\mathbf{T}}(\tau)]^k$$

3. Now predict $\mathbf{p}(k\tau)$ from the same initial state using the $\widetilde{\mathbf{T}}(k\tau)$ matrices:

$$\mathbf{p}^\top(k\tau)_{\mathrm{Trajectory}} = \mathbf{p}(0)^\top \widetilde{\mathbf{T}}(k\tau)$$

4. Compare the $\mathbf{p}(k\tau)_{\mathrm{HMM}}$ and $\mathbf{p}(k\tau)_{\mathrm{Trajectory}}$ for a range of values of k and for different initial state vectors, $\mathbf{p}(0)$. In PyEMMA (version 2.5) [253] $\mathbf{p}(0)$ are the HMM basis vectors, i.e. $\mathbf{p} = (1, 0, 0, \ldots),\ (0, 1, 0, \ldots)$.

2.6 Summary

This chapter has described the theory underlying constructing Markov state models in a fine grained microstate basis and then coarse-graining this model using hidden Markov models. To summarise the methods detailed here, starting from a set of molecular dynamics trajectories. *Step 1*: choose a feature, χ, related to the slow dynamic processes being studied and project the atomic coordinates of the trajectories onto this feature. *Step 2*: perform a time-lagged independent component analysis (TICA) parameterized with a lag time of τ', and project the feature trajectories onto the first m TICA components. This step is optional and is not used in the work of chapter 3 which estimates Markov models of water dynamics in aerosol particles. *Step 3*: cluster the reduced dimension trajectories

into n microstates used k-means clustering. *Step 4*: determine an appropriate Markov lag time by estimating Markov state models at different lag times and looking for the smallest lag such that the slow implied timescales remain constant. *Step 5*: estimate a Markov state model at this lag time, τ, and determine the number of dominant processes, r, by looking for gaps in eigenvalue spectrum of $\mathbf{T}(\tau)$. This model is then specified by the hyperparameters (χ, τ', m, n). *Step 6*: Score this model by calculating the cross-validated VAMP-2 score using the r dominant eigenvectors. *Step 7*: repeat steps $1-3$ varying the values of the hyperparameters and re-scoring the resulting models, while keeping the number of dominant eigenvectors, r, and the Markov lag time, τ, fixed. Then choose the set of hyperparameters which maximize the VAMP-2 score. A more efficient method for performing these steps, utilising ideas from the machine learning community, is investigated in chapter 4 using the benchmark system of alanine dipeptide. *Step 8*: coarse-grain the optimum MSM using a hidden Markov model with r hidden states representing the r metastable states, as determined in step 5. In chapter 5 a different method for determining the number of metastable states is investigated. This method uses more abstract classification techniques taken from the statistics community and tests them on a model four-well system. Chapter 6 incorporates the work of chapters 4 and 5 into this general method in order to develop a Markov model description of the conformational dynamics of the enzyme aromatic amine dehydrogenase.

CHAPTER 3

TRANSIENT CAVITY DYNAMICS AND DIVERGENCE FROM THE STOKES–EINSTEIN EQUATION IN ORGANIC AEROSOL

Contributions

This chapter contains an adaption of "Transient cavity dynamics and divergence from the Stokes–Einstein equation in organic aerosol" by Young-Chul Song, Stephen Ingram, Robert E. Arbon, David O. Topping, David R. Glowacki, and Jonathan P. Reid. This was published in the journal Chemical Science, volume 11, pages 2999-3006.

Changes have been made including figures, section and reference numbers to suit a thesis structure and maintain formatting conventions with the rest of this thesis. Additional discussion of the results and conclusions have also been added. Some of the supplementary information from the paper has been included in the main text, the remaining supplementary information can be found in appendix 8.

Contributions to the work: Young-Chul Song performed the experimental work; David O. Topping performed the analysis of diffusion constants from the experimental results; Stephen Ingram performed the MD simulations, the determination of the diffusion constants from MD simulations, the exploration of the cavity dynamics and the packing efficiency. The author of this thesis contributed to sections 3.4, 3.1 and 3.5 of this work. Specifically:

1. Suggested the splitting of the molecular dynamics trajectories into times-slices, although did not do the analysis of the sucrose cavities (figure 3.3 and 3.5).
2. Performed all Markov state modelling (exemplified in figures 3.4 and 3.6)) and wrote the Markov state modelling section of the supplementary material of the published paper, which

has been incorporated into this chapter.

3. Classified the time-slices as being in equilibrium and non-equilibrium, and calculated the water hopping barrier heights (figure 3.7) for the equilibrium time-slices.
4. Added discussion of Markov state models to the introduction (section 3.1) and of the Markov analysis to the conclusions (section 3.5).

The work was supervised by David R. Glowacki and Jonathan P. Reid.

3.1 Introduction

Examining the relationship between the diffusion rates of small molecules and the viscosity of the surrounding molecular matrix is important for exploring problems as diverse as the molecular mechanisms of crystallization and the formation of amorphous phases in drying droplets [260–262], the controlled-release of active ingredients from structured micro-particles in pharmaceutical and consumer products [263–266], and the mass concentration of secondary organic aerosol particles in a polluted urban environment [267–269]. The simplest relationship, the Stokes–Einstein (S–E) equation, expresses the inverse correlation between the translational diffusion coefficient, D, of a large spherical solute molecule of radius a, moving within a solvent continuum with a dynamic viscosity, η [120, 270]:

$$D = \frac{k_B T}{C \pi \eta a} \tag{3.1}$$

where C is a constant. However, in many important cases the "solvent" (i.e. the dominant component by mole fraction) may be a large organic molecule and the "solute" (i.e. the minor component) may be a small molecule, e.g. water [120–122]. For example, in the drying of aqueous-organic solution droplets, the evaporation of water can lead to an involatile solute surpassing its solubility limit, thereby becoming the major component with a mole fraction that can approach 1. The sudden removal of water can lead to a "frozen" organic-rich matrix with a sufficiently high viscosity such that nucleation and crystallization are delayed, unable to occur on an experimentally realisable timescale, with the solution composition crossing the threshold for a moisture-induced glass transition [129]. Even then, the residual moisture content can impact product lifetime and particle morphology. Under these conditions, it is most appropriate to consider the diffusion of water within an organic matrix at infinite dilution of water; however, it is typical that a significant divergence from the S–E equation is observed in this limit [120, 123, 124]. Modifications to the S–E equation have been suggested, including the use of a fractional exponent (i.e. $D \propto \eta^{-\alpha}$), that account for different relationships between the diffusion coefficient and viscosity [125–127].

Independent measurements of diffusion coefficients and viscosities over the appropriately wide ranges needed to observe the failure of the S–E equation are challenging. Most measurements report the temperature-dependence of viscosities and diffusion coefficients for super-cooled liquids or solutions of fixed composition, and can approach close to the glass transition temperature

[270–273]. By contrast, there are many fewer studies of the compositional dependence of the divergence from the S–E equation, for example with diminishing moisture content as the glass transition relative humidity (RH) is approached [120, 121, 123, 124, 126, 274]. Moisture acts as a plasticizer in atmospheric aerosol particles, regulates the viscosity and, thus, shelf-life of amorphous particles used in formulations, and could play a critical influence in governing crystal formation in drying droplets and films as opposed to the formation of an amorphous solid. Examining the compositionally dependent divergence of an organic solute–water mixture from S–E behaviour not only requires accurate measurements of diffusion coefficients and viscosities over as much as 15 orders of magnitude but requires accurate measurements of composition, recognising that both viscosity and diffusion coefficients are highly dependent on the identity for the functional groups forming the organic solute [275]. To access the full viscosity range, moisture must be removed from metastable supersaturated solution droplets without crystallization.

Reported here is a systematic experimental and computational study of the failure of the S–E equation for a range of aqueous-saccharide solutions, varying the molecular size of the organic molecule forming the viscous matrix relative to water and exploring the detailed mechanism of water transport in the limit of a pure saccharide particle. The experimental measurements are complemented with $9\,\mu s$ of molecular dynamics (MD) simulations at atomistic resolution for a single type of saccharide matrix. In order to understand the microscopic dynamics of the water transport a Markov model (MM) approach was used. Much of the recent work [60, 130] in Markov modelling has been on the feature selection problem: creating variationally optimised basis sets which capture the slow dynamics of the system. However, this work uses the Cartesian coordinates of the water molecule, with no variational optimisation, as a method for both i) partitioning and classifying the dynamics as being in either local equilibrium or non-equilibrium and, ii) determining the typical free energy barriers faced by water as it moves through the saccharide matrix. This chapter is structured as follows. In section 3.2 the experimental measurements of diffusion coefficients are described and discussed; section 3.3 makes the link between the diffusion and viscosity; section 3.4 describes the elucidation of the microscopic mechanism from molecular dynamics simulations and MMs; section 3.5 concludes and discusses limitations of the Markov modelling.

3.2 Measurements of diffusion coefficients of water in aqueous-saccharide aerosol particles

Not only are saccharides used widely as excipients for drug delivery [276–278] and excipient particles are often prepared by spray drying [279–281], they find widespread application in the food industry and are commonly used as laboratory surrogates for high oxidized viscous secondary organic atmospheric aerosol [129, 262, 282–285]. Using aerosol particles levitated in optical tweezers, measurements were carried out which avoid the process of heterogeneous nucleation that occurs in the presence of a substrate, allowing access to particle viscosities spanning dilute aqueous solutions ($10\,\mathrm{mPa\,s}$)

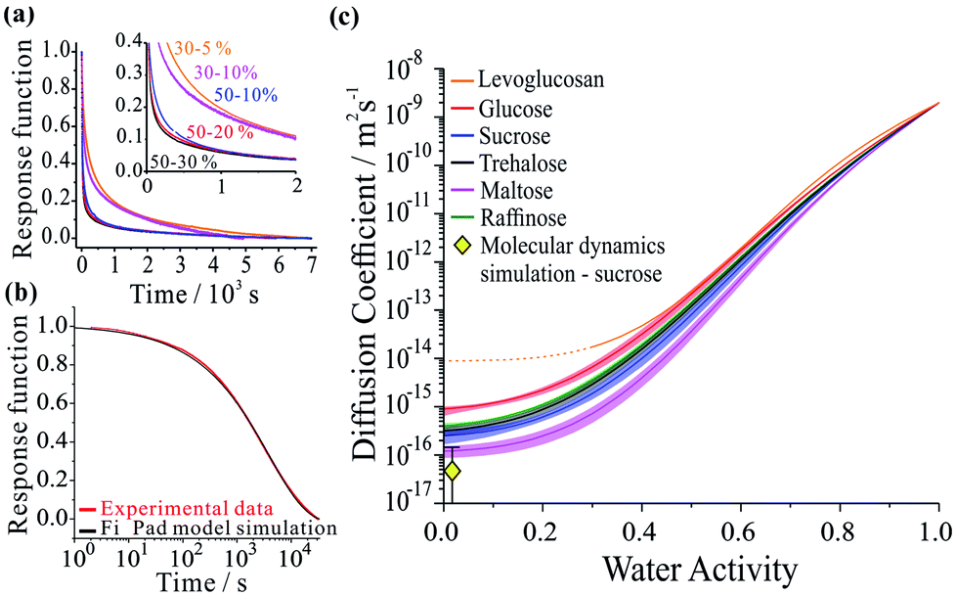


FIGURE 3.1: EXAMPLE EXPERIMENTAL WORKFLOWS. Examples of each step in the workflow required to extract the compositional dependencies the diffusion coefficients from a time-dependence in particle size. The panels show: (a) a collection of response functions for size changes of aqueous-raffinose particles following a step change in RH; (b) a single response function following a change in RH from 30 to 5 RH for a sucrose particle and the best-fit produced by the Fickian diffusion model. (c) The estimated compositional dependencies of the diffusion of water in the six binary aqueous-organic aerosol systems studied. The estimate of the diffusion coefficient for water in sucrose from the MD simulations also presented (yellow diamond).

to an amorphous solid (10 TPas). The moisture content is readily altered by varying the relative humidity of the gas phase. Specifically, considered here are five binary aqueous-saccharide solution aerosols: glucose (a mono-saccharide); sucrose, trehalose and maltose (all di-saccharides); and raffinose (a tri-saccharide). Also considered are aqueous aerosol droplets containing levoglucosan, a representative oxygenated compound of biomass burning aerosol particles in the atmosphere [286].

Figure 3.1 panel (a) shows examples of the time-dependence in the size response functions for aqueous-raffinose particles following transitions in RH. The significant changes in the equilibration time reflect the significant changes in particle viscosity that are observed over this range in RH/moisture content: equilibration at RHs above the glass-transition RH occurs on timescales $\ll 1\,\mathrm{h}$; at low RH, the release of moisture from an amorphous glass occurs over many hours and indeed is not complete over the experimental timescales. The time-constant, τ, and "stretch factor" β of the multiexponential decay observed in both evaporation and condensation events show a path dependence, varying with both the initial and final RH, the initial particle size and the wait-time at intermediate RHs (see figure 8.4). Tabulated values of both parameters observed in each of the new systems may be found in table 8.1. To fit the compositional/water activity dependence of the diffusion coefficient of water for binary solution aerosol droplets requires measurements at many RH

transitions [111, 287]. Measurements were performed over 6 RH transitions for 96 droplets for the six binary aqueous-organic aerosol systems studied (glucose, sucrose, trehalose, maltose, raffinose and levoglucosan). Transitions in size were slowest for maltose droplets at the lowest RHs. Moreover, the characteristic timescale increases with increasing particle size for every binary organic system studied (see figure 8.3 and table 8.1). Time-constants for all particle sizes in the range $3\,\mu m$ to $6\,\mu m$ show the same ordering: maltose > raffinose > trehalose > sucrose > glucose ≥ levoglucosan. In other words, the chain length of the organic fraction appears to be important to the internal mixing dynamics but is not the only controlling factor.

The compositional dependencies of the diffusion coefficients of water estimated for these binary aqueous-organic systems are summarized in panel (c). For reference, the moisture driven glass transition RH has been reported as 53 % for raffinose [285, 288] and 32 % for maltose [285]; the majority of evaporation measurements for these two systems have been made with ultra-viscous and even glassy particles. A value of 23 % RH has been reported for sucrose [285, 288] while glucose and levoglucosan are not expected to become glassy at any moisture content at this temperature [283]; indeed, levoglucosan crystalizes at an RH of 30 % and diffusion coefficients cannot be measured below this. The trend in D_w is not monotonic with molecular weight: levoglucosan ($162.1\,g\,mol^{-1}$) > glucose ($180.2\,g\,mol^{-1}$) > raffinose ($504.4\,g\,mol^{-1}$) > trehalose ($342.3\,g\,mol^{-1}$) > sucrose ($342.3\,g\,mol^{-1}$) > maltose ($342.3\,g\,mol^{-1}$) at the same water activity. Water in the monosaccharide shows the fastest diffusivity, and diffusion in the trisaccharide is faster than in the disaccharides when a fixed RH/water activity is considered. Indeed, this trend in the diffusion coefficient of water in the limit of a pure dry organic matrix is consistent with a previous assessment of the diffusion coefficients at the glass transition temperature for a subset of the compounds studied here [289].

3.3 The relationship between diffusion and viscosity in mono-, di- and tri-saccharide particles

The diffusion coefficient measurements presented in figure 3.1 and our measurements of solution droplet viscosities [285] allow us to examine their correlation over wide ranges spanning more than 12 orders of magnitude in viscosity and 7 orders of magnitude in diffusion coefficient. The correlations for these systems are compared with predictions from the S–E equation in figure 3.2, assuming a molecular diameter for water of $0.2\,nm$. Typical error estimates in diffusion coefficient and viscosity are indicated by the representative error bars for each system. The diffusion coefficients for water in all organic-aqueous solutions increasingly deviate from the S–E equation with decreasing water activity and increasing viscosity. Even at the threshold of semi-solid behaviour ($104\,Pa\,s$), the diffusion coefficient of water in aqueous-raffinose aerosol droplets is ~5 orders of magnitude larger than estimated by S–E. This is a consequence of the inapplicability of the S–E assumptions to estimations of the diffusion coefficient of a small molecule moving within a matrix of large molecules, i.e. the translation of water is not characterized by simple Brownian motion [272, 290].

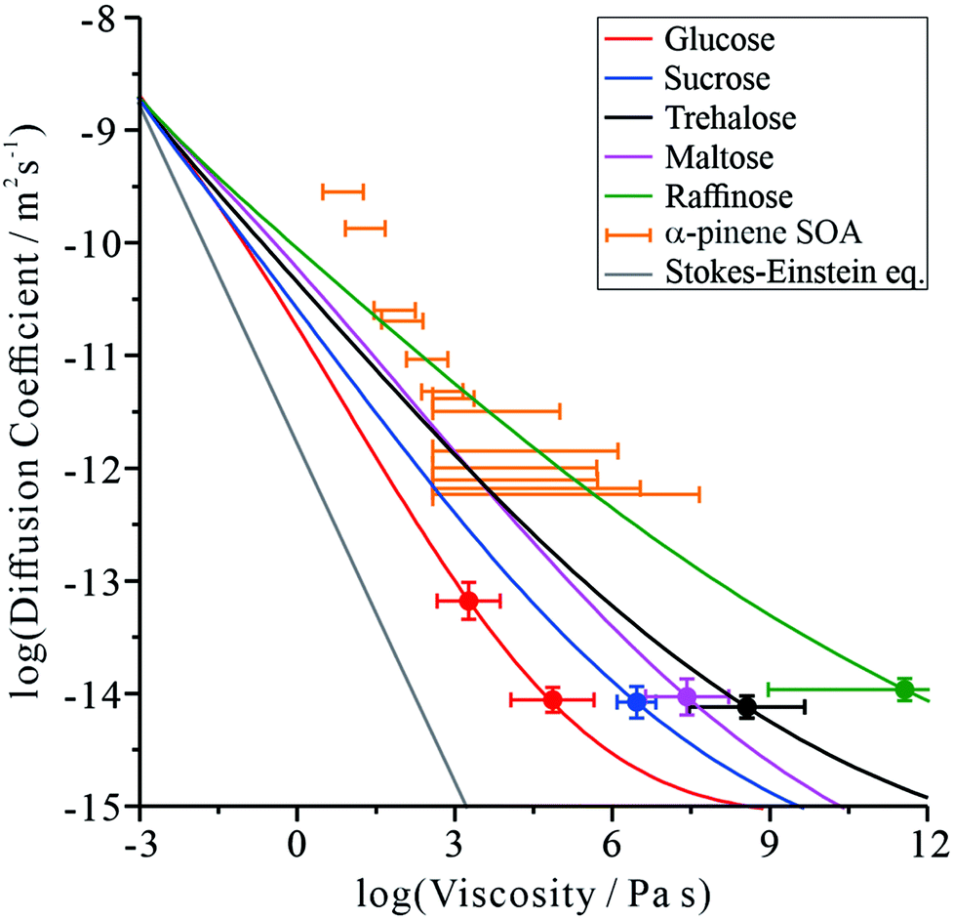


FIGURE 3.2: CORRELATION OF THE DIFFUSION COEFFICIENT OF WATER WITH THE VISCOSITY OF THE AQUEOUS-ORGANIC MATRIX. A prediction from the S–E equation is shown by the grey line. The relationship between the diffusion constant and viscosity of α-pinene SOA (orange markers) has been inferred by us from the literature, and has previously been discussed.[123, 291] The colour scale is the same as in figure 3.1.

Comparing the relative divergence of water diffusion coefficients from S–E predictions for the mono-, di- and tri-saccharides, the discrepancy increases systematically across this series. Water transport is fastest in solutions with the tri-saccharide raffinose and slowest in solutions with the mono-saccharide glucose at a certain solution viscosity; the di-saccharides (sucrose, trehalose and maltose) fall in the intermediate range. These results suggest that the disparity in size between water and the organic molecule forming the matrix is key to determining the diffusion rate of water. It also explains why the particle size relaxation times and limiting D values (in dry air) did not directly scale with molecular weight: the particles exhibit different viscosities at the same water activity. Therefore, an independent viscosity axis, in this case produced using the aerosol particle coalescence technique, is crucial to separating the two effects.

It can be postulated that when forming a matrix from raffinose, a much larger molecule than

water, the packing density of raffinose leaves sufficient free volume for water to move more readily through the network of organic molecules. When the organic molecule is closer in size to water, as in the case of the mono-saccharide glucose, the tighter relative packing of glucose leads to a fewer adequately sized cavities. In this sense, the mechanism of impaired water transport more closely resembles percolation rather than diffusion [292, 293], a process that is sensitive to the free volume of the medium [294].

Figure 3.2 is instructive when considering the diffusion of water through the complex organic matrices found in atmospheric secondary organic aerosol (SOA), one particular motivation for the current study. For example, water transport in α-pinene SOA (orange bars) is more rapid than would be expected based on measurements of viscosity and estimates from the S–E equation [120, 122, 123], most closely resembling the di- and tri-saccharides. However, it should be recognised that the properties of SOA constituent molecules are considerably different. The average molecular weight of organic components identified in α-pinene SOA is $150\,\mathrm{g\,mol^{-1}}$ to $200\,\mathrm{g\,mol^{-1}}$ [291], albeit with a lower degree of oxygenation: typically the O:C ratio has been reported as 0.45 to 0.55 [295]. The O:C ratios for trehalose and raffinose are 0.92 and 0.89, respectively. The faster diffusion of water in SOA than expected from the S–E equation may be attributed to the heterogeneity in composition at the molecular scale, leading to a porous network of channels through which water transport is more facile than expected.

3.4 The microscopic mechanism from molecular dynamics simulations

To better understand the microscopic mechanism of water transport, atomistic molecular dynamics (MD) simulations of water in sucrose (with 'concentrations' of one water molecule per 35 sucrose molecules) were carried out which were designed to mimic experimental water activities close to zero. See appendix 8 for further information on the MD simulations. The initial placement of the organic molecules is intended to replicate the amorphous packing structure that is expected to occur near the surface of a glassy sucrose droplet [261]. These MD simulations were analysed to provide an independent estimate for the value of the intercept $D_{\mathrm{w,org}}$. Figure 3.1 panel (c) shows the MD-derived value of $4.64 \times 10^{-17}\,\mathrm{m^2\,s^{-1}}$ is in good agreement with the experimental measurements, and indicates that our computational approach captures the physics of water diffusion in sucrose at low activities.

Inspection of our MD results reveals that the mechanism of water diffusion through the sucrose matrix proceeds by a hopping between cavities (figure 3.3). In general, a 'cavity' is defined as a sucrose interstitial domain where water has a significant lifetime based on Markov analysis. Our analysis has enabled us to identify both reversible and irreversible examples of intercavity dynamics. Figure 3.4 and much of the analysis described in this article focuses on clusters of 'cavities' between which water molecules make reversible kinetic hops, because this local equilibrium is amenable to analysis using standard tools in statistical mechanics.

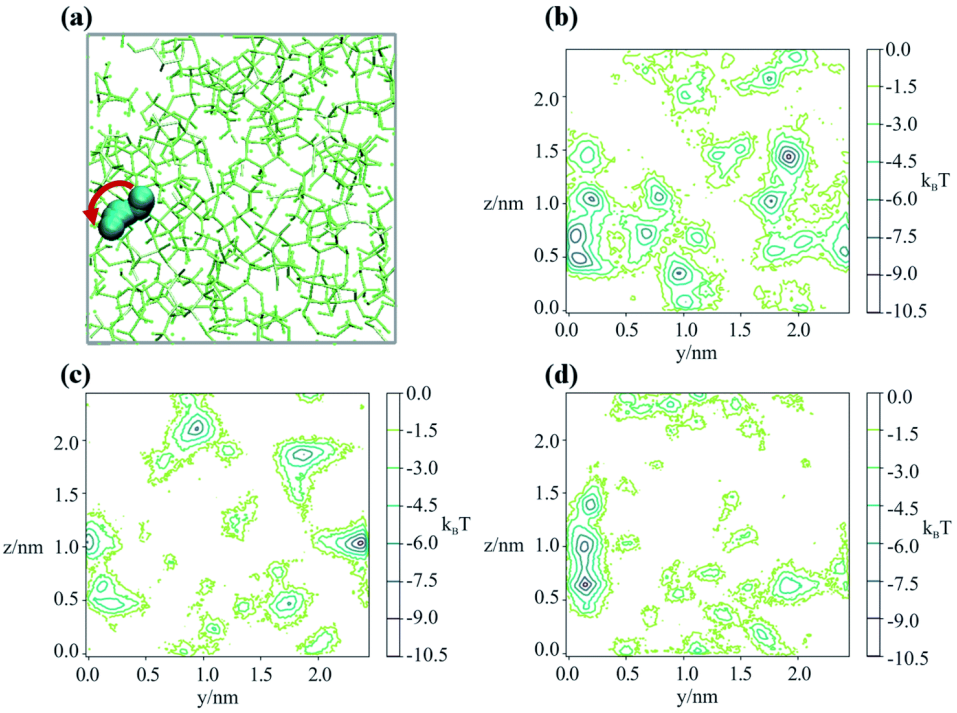


FIGURE 3.3: SUCROSE MATRIX CAVITIES. (a) Snapshots in the yz plane of the trajectory of a water molecule 'jumping' between interstices in an amorphous sucrose lattice. The red arrow indicated the direction of the observed hop. The periodic box is shown in grey. Sequential calculated potentials of mean force that the water experiences are shown in panels (b) to (d), separated in time by 1 ns.

Our analysis shows that water remains in a cavity (or cluster of cavities) until either (1) it achieves sufficient kinetic energy to escape the local environment, or (2) the slower dynamics of the sucrose matrix opens a pathway that allows access to a new cavity. This appears similar to the 'micropore diffusion' mechanism, which has been proposed to describe the uptake and transport of small molecules through porous zeolite structures [296].

In order to determine the time-dependent dynamics of the cavities, a 3 ns timeslice of a 1 ns trajectory was identified, where a water molecule jumping between two distinct cavities was observed, as illustrated in figure 3.3 panel (a). Over the course of the 3 ns timeslices, three different equilibrium configurations were extracted, and a 50 ns MD simulation beginning from each of these points was run, freezing the sucrose but not the water. The purpose of these simulations was to use the water molecular as a "probe" of the cavity structure and dynamics, in order to understand cavity persistence on the timescale of a typical water hop. The potential of mean force within each cavity (PMF), without the entropic degrees of freedom of the organic matrix included, was determined by Boltzmann weighting the resultant probability distribution, P:

$$\mathrm{PMF} = -k_\mathrm{B} T \ln(P) \tag{3.2}$$

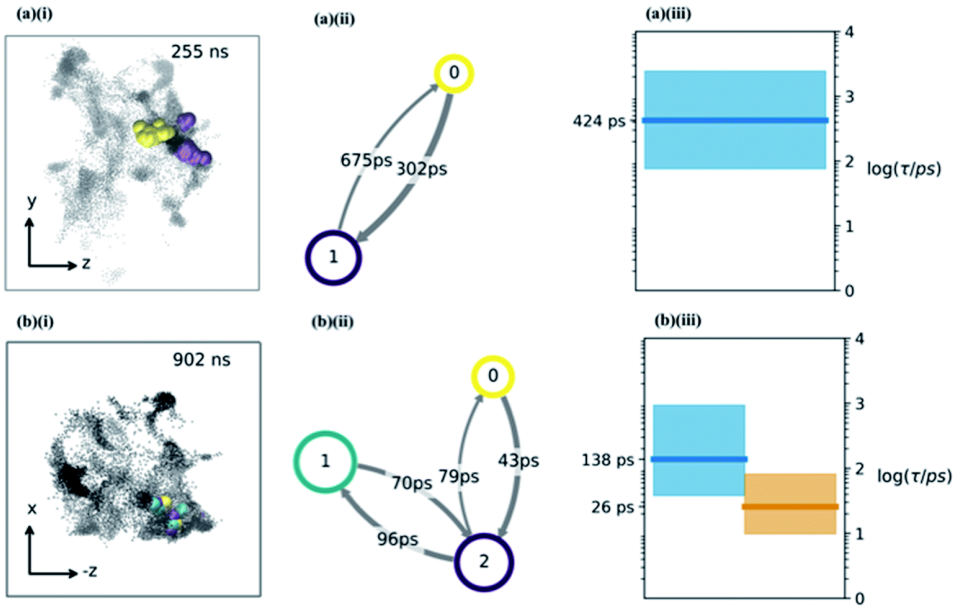


FIGURE 3.4: THE WATER HOPPING MECHANISM. (a and b) The water hopping mechanism showing 2 and 3 metastable states (panels (a) & (b) respectively) arising from one of the nine trajectories. Subplots (i) show zy, (-z)x projections of water molecule's position throughout the trajectory with different colours indicating different metastable states. Subplots (ii) shows a hidden Markov state model representation of the hopping behaviour. Each circle represents a metastable state, with the size related to its stability, the arrows show the hopping timescale in picoseconds from one state to another (same colour scheme as (i) subplots). Subplots (iii) show Bayesian estimates of the relaxation timescales associated with hopping between the states (thick line is mean, coloured region is a 95 % credibility interval).

Figure 3.3 panels (b) to (d) shows that there is a small but noticeable change in the cavity PMF landscape (the region around $y = 0.1$, $z = 0.5$–1) as the sucrose reorients over 3 ns. This observation is consistent with analysis showing that the position–position autocorrelation function of a single sucrose molecule decorrelates after approximately 1 ns, as presented in figure 3.5. Having determined an approximate upper limit on the sucrose re-organisation timescale, each trajectory was split up into 1 ns slices and the kinetic parameters for water-hopping between cavities were determined using a Bayesian Hidden Markov (HM) modelling approach.

The sucrose reorganisation time of 1 ns defines a time scale over which the molecular environment, which defines the free energy surface over which the water moves, remains stationary. Over this time scale the rates of water diffusion can be assumed to be independent of time, which admits a Markov analysis (where the transition rates do not vary with time). Within these stationary time-slices, the water molecule was typically observed to remain trapped within a small region made up of one or more cavities as shown in figure 3.3. The dynamics of the water molecule can be classified as either reversible or irreversible with respect to these small regions. If the dynamics were reversible then the absolute probability of observing hops from one part, A, of the region to another part B, of the

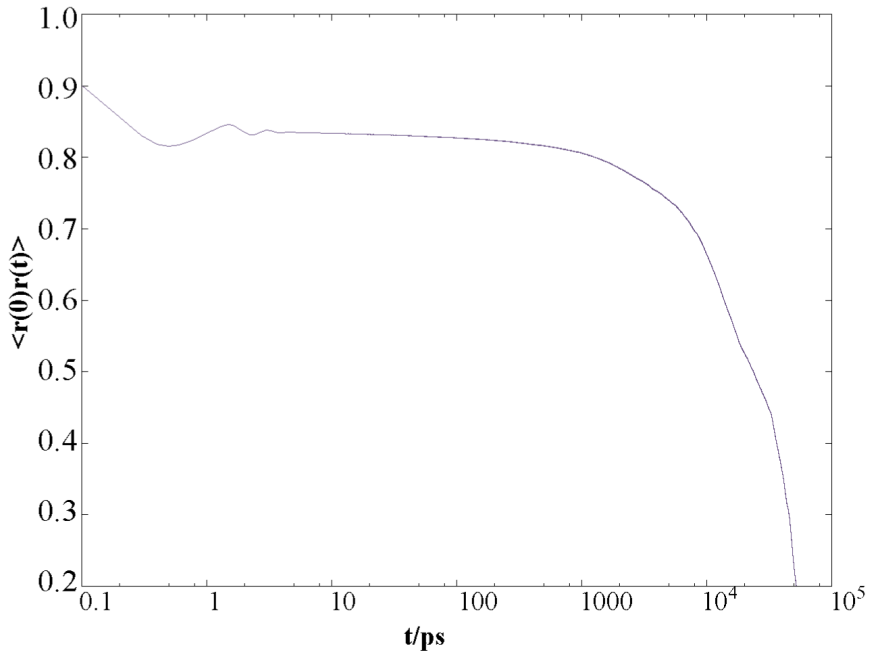


FIGURE 3.5: POSITION-POSITION AUTOCORRELATION FUNCTION OF A SINGLE SUCROSE MOLECULE WITHIN THE SIMULATION. The plot shows very little motion in the region 0.1 ps to 1000 ps.This plot suggests that it is reasonable cluster the data (to determine cavities, as in figure 3.4) in increments of 1 ns. It is also reasonable to initialise frozen sucrose simulations after intervals of 1 ns (figure 3.3, panels (b) - (d))

region (between two cavities within the region, or within a single cavity) is the same $A \rightarrow B$ and $B \rightarrow A$. In other words, detailed balance is observed when concentrating on only one time-slice. If this criteria is not fulfilled, the dynamics were said to be irreversible.

The goal of the Markov modelling was to identify regions of the MD simulations in local equilibrium,to identify cavities,and calculate their associated hopping and relaxation timescales. Eight MD simulations were partitioned into 1 ns time-slices and the position of the water center of mass was clustered into 100 discrete states using the k-means [244] clustering algorithm. A Markov lag time of $\tau_M = 10\,\text{ps}$ was used based on the implied timescales of 10 randomly selected time-slices from trajectory 3. The k-means clustering and all subsequent calculations were performed using the open source software, PyEMMA (version 2.5) [253].

The discrete trajectories were screened to see whether they were in local equilibrium with the following procedure:

1. A Markov state model (MSM) with a lag time (τ_M) of 10 ps was constructed for each time-slice. The lag time was chosen so as to reveal details of potential metastability [64]. To determine the value of τ_M a sensitivity analysis was carried out, varying τ_M until convergence was observed in the timescales. See figure 3.6 panel (b) & (f).

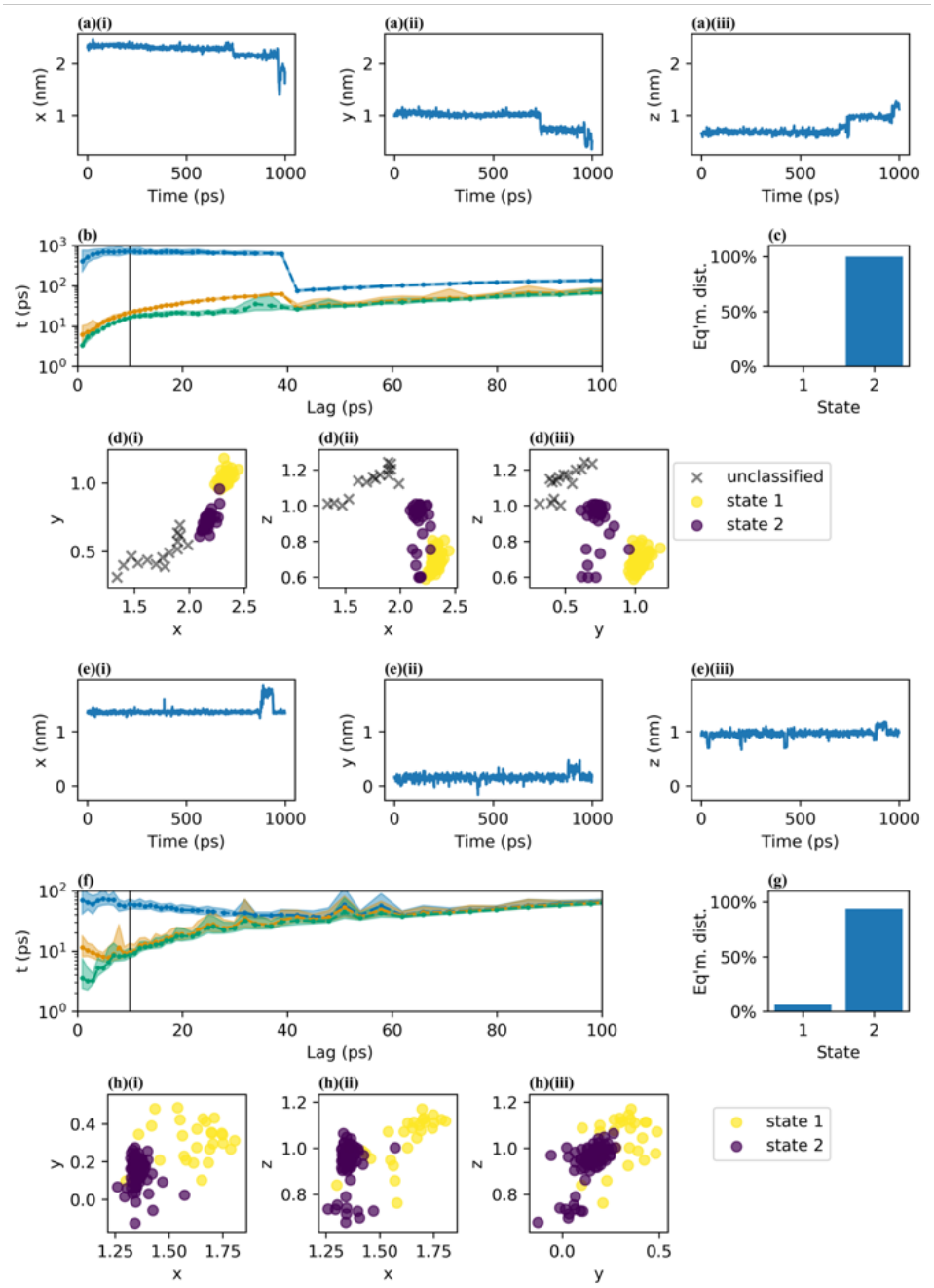


FIGURE 3.6: CLASSIFICATION OF TWO TIME-SLICES FROM TRAJECTORY 3: 0 ns to 1 ns (panels (a) - (d), non-equilibrium time-slice) and 22 ns to 23 ns (panels (e) - (h), local equilibrium time-slice). Panel (a) and (e) show the x, y, and z coordinates of the center of mass of the water molecule. Panel (b) and (f) show the sensitivity analysis to determine the lag time τ_M: the implied timescales, (t vertical axis) vs the MSM lag, (τ_M horizontal axis). The black vertical line is placed at 10 ps is the minimum time at which the implied timescales show convergence. Panels (c) and (g) show the stationary distribution. In panel (c) state 1 has negligible probability classifying this time-slice as not being in equilibrium. Panels (d) and (h) show the attempted assignment of the discrete trajectory into metastable states. The non-equilibrium time-slice in panel (d) shows the large portion of unclassified states.

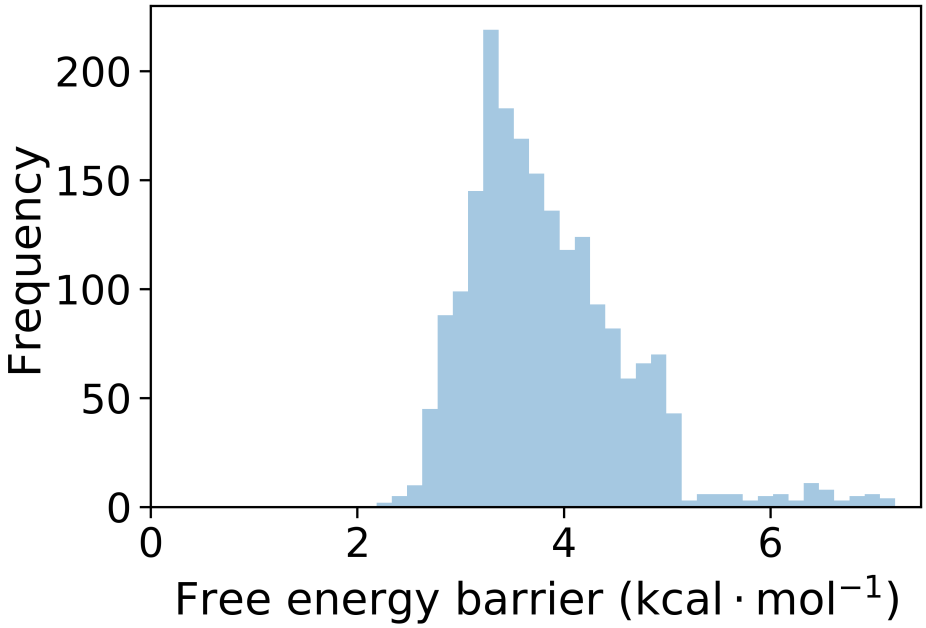


FIGURE 3.7: HISTOGRAM OF BARRIER HEIGHTS FOR EIGHT TRAJECTORIES. Barrier heights were calculated in 1 ns intervals. Free energies are calculated by converting the hopping timescales (t) of the hidden Markov models to free energies (ΔG) using the transition-state theory expression for the hopping rate $k = 1/t$: $\Delta G^{\mathrm{TS}} = RT\ln(k_b T \cdot t/h)$.

2. This MSM was coarse grained into a k-state Hidden Markov Model (HMM) if the largest gap in successive implied timescales of the MSM (t_k / t_{k+1}) was greater than 1.5.
3. If the HMM had:
   a) an absorbing state (a 1 on the diagonal of the transition matrix);
   b) an metastable state, i, with stationary distribution (π_i) probability approximately zero ($\pi_i < 10^{-8}$)
   c) negative eigenvalues;

   it was classified as not being in local equilibrium.The conditions (a) – (c) are indicative of not being in equilibrium because of the enforcement of the detailed balance conditions when estimating the HMM transition matrix [169].
4. The time-slices identified as being in local equilibrium were used to construct a Bayesian HMM in order to estimate errors of the hopping timescales.

Out of the 8000 time-slices analysed 947 (11.8 %) were positively identified as being in local equilibrium. Figure 3.6 demonstrates this classification algorithm for two time-slices of trajectory 3. The first time-slice (0 ns to 1 ns, panels (a) to (d)) was classified as not being in local equilibrium, the second time-slice (22 ns to 23 ns, panels (e) - (h)) was classified as being in local equilibrium.

For each of the 947 1 ns slices in local equilibrium, a HM model was constructed to determine

the number of metastable states (cavities), the relative population of water within each cavity, transition probabilities for hopping between cavities, and the timescales for inter-cavity transport. The HM analysis shows that the cavities have a distribution of free energies (figure 3.7), and a corresponding distribution of lifetimes for water within any given cavity. Figure 3.4, panels (a) and (b) show representative examples from three $1\,\mathrm{ns}$ slices throughout one MD simulation where water hops between clusters of two and three cavities, along with information regarding the hopping timescales in subplots (ii). Subplots (iii) show the timescales of interstate rearrangement processes. These timescales do not correspond to pairwise hopping between cavities but rather joint relaxation processes over all states.

Extended periods of cavity hopping behaviour are found in all our $300\,\mathrm{K}$ MD simulations: water repeatedly moves back and forth between adjacent cavities that do not fully collapse once they are vacated. The characteristic barriers for hopping between cavities have been calculated using transition state theory and are found to be on average $6.42 \pm 1.29 k_b T$ $(3.83 \pm 0.77\,\mathrm{kcal\,mol^{-1}})$. The distribution across all trajectories (figure 3.7) corresponds to a hop frequency of between 1 and 50 per nanosecond per water molecule, although not all hops will lead to productive diffusion against a concentration gradient. In fact, 'return trips' may be a common feature of water transport in these matrices. Our analysis suggests that the magnitude of the S–E deviation depends on the transient packing efficiency of the organic molecules. For instance, raffinose self-diffuses slower than sucrose (hence the observed particle viscosity is higher), but if the average volume of cavity space within the lattice is larger and more highly connected, then the net water flux will be higher at a given particle viscosity. To evaluate this, a series of MD simulations were carried out, which were post-analyzed to assess the cavity volume within glucose, sucrose and raffinose matrices. The final coordinates of the three matrices are presented in figure 3.8, showing increasing cavity size and density within the van der Waals surfaces. This quantity is also expressed as a fraction of the simulation volume in figure 3.8.

Thus, the trend in viscosity data (figure 3.2) can be rationalised at a molecular level: the 'hopping' mechanism of water transport will become more efficient as the size of the organic constituent increases. Therefore, the frictional forces experienced by water molecules will deviate further from those assumed by equation 3.1, and the observed D will be under-predicted to a greater extent for larger organics. With reference to atmospheric organic aerosol, this effect may be significant in particles containing large numbers of oligomeric or 'humic-like' molecules. Such constituents are frequently found in aerosol formed under low RH [297], low temperature [298] or high precursor concentration [299] conditions.

3.5 Conclusions

In conclusion, this work has shown that the diffusion constant of water in viscous aerosol particles departs increasingly from the S-E equation as the size of the saccharide molecule forming the matrix

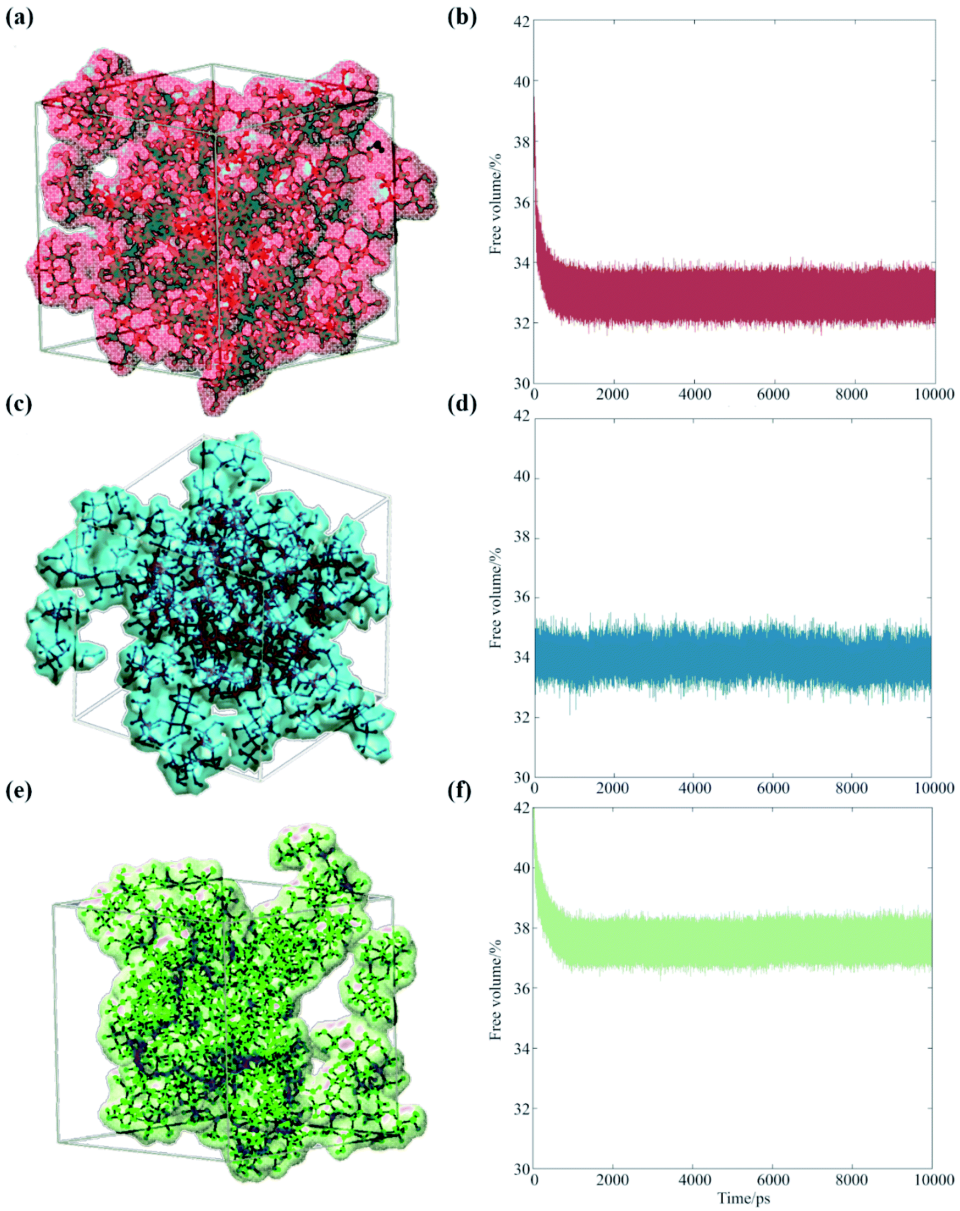


FIGURE 3.8: PACKING EFFICIENCY OF ORGANIC MOLECULES. Occupied volumes of (a) glucose (67.1 %), (c) sucrose (66.0 %) and (e) raffinose (62.4 %) matrices are shown (same colour scheme as figure 3.1 again) showing the van der Waals radii of the saccharides within one snapshot of the short simulations. Fractional free volume are shown for (b) glucose, (d) sucrose, and (f) raffinose calculated by GROMACS throughout 10 ns simulations.

increases. Atomistic simulations suggest that larger molecules will pack less efficiently, facilitating a mechanism of activated hopping through the porous network: at high saccharide fraction, a water molecule executes discrete jumps between cavities at a rate governed by the collective motion of the saccharide matrix. These observations also are consistent with the slower diffusion of molecules larger than water, whose motion more closely resembles that described by Stokes flow [273].

The Markov modelling approach adopted here facilitated both the i) decomposition of a large amount of simulation data into tractable time-slices with an approximately stationary potential induced by the sucrose matrix; and ii) the quantitative description of kinetic processes involved in each slice. The results of applying the equilibrium assumptions to classify the time-slices resulted in only $11.8\,\%$ being positively identified as being in local equilibrium.

A number of assumptions and approximations were made in order to facilitate this analysis. The first assumption was that the water dynamics could be approximated by considering a water molecule traversing a free energy surface arising from an static sucrose environment. This was was justified on the separation of timescales between the sucrose and water molecules. The estimated reorganisation time of $1\,\mathrm{ns}$, over which this assumption was assumed to hold, pertains to the average movement of the sucrose molecule. However, fluctuations in parts of the sucrose molecule on the same timescale of the water motion were not ruled out. If present, these motions break the assumption that the kinetic rates of the water molecule were independent of time. These interactions were also left out of the variables used to describe kinetic states of the water molecule which limits the accuracy of eigenvectors of the Markov state model, i.e., some "essential" degrees of freedom were missing from the kinetic description.

The second assumption was that, even if more rapid fluctuations of the sucrose molecule could be ruled out, the assumption of a universal reorganisation timescale of $1\,\mathrm{ns}$ was likely inaccurate. This results from two factors. First, this timescale was an average and thus deviation from this value are expected. Second, even if deviations from an average reorganisation timescale are minimal, this particular figure is likely inaccurate because the sucrose reorganisation time was measured on a single sucrose molecule from a single trajectory by inspection of the autocorrelation function. As a result, the proportion of time-slices described as being in equilibrium underestimates the true proportion of local equilibria. This is because it does not rule out local equilibria on time-slices shorter than $1\,\mathrm{ns}$ which may have then been incorrectly classified as out of equilibrium. Similarly, local equilibria may have lasted longer than $1\,\mathrm{ns}$ which would have allowed combining observations to produce a more precise estimate of the hopping timescales.

The third assumption concerns the equilibration of the simulations. Each trajectory was equilibrated for $500\,\mathrm{ps}$ and measurement of the diffusion constant of water in the subsequent microsecond of simulation produced a value in agreement with the experimental measurements. However, by their nature substances near the glass transition state have long equilibration times [129]. It is possible that the rate of cavity formation and their size and shape could have changed over the course of the trajectories as the simulations continued to equilibrate after the initial $500\,\mathrm{ps}$. Previous

assumptions notwithstanding, this would not affect the validity of the Markov analysis because the trajectories were partitioned into stationary time-slices. However it may affect the distribution of hopping barriers and the measurement of diffusion constant.

In addition, the modelling approach misses certain dynamical processes. First, it does not say anything about the non-equilibrium processes which form the majority of the water transport. Second, the analysis of the equilibrium time-slices focused only on the slowest timescale processes in each time-slice. Multiple gaps in the eigenvalue spectrum were not identified which could have revealed further, faster timescale processes. A plan to address these limitations in future work is laid out in the conclusions, chapter 7.

Despite this, the simplified Markov modelling approach used here, provided an insightful explanation of the diffusion of water within the saccharide matrix, in-line with experimental measurements and other simulation analysis. The models produced were successfully validated and visualisation of models from individual time-slices appeared consistent with the simulation data. The simplified approach used intuition, visualisation and heuristics from the literature to guide the modelling process. The "essential degrees of freedom" - the Cartesian coordinates of the water molecule - were suggested by inspection of the molecular dynamics trajectories. The number of microstates followed a simple heuristic related to the volume of data, while the number of metastable states was determined from the eigenvalue spectrum.

CHAPTER 4

MARKOV STATE MODEL OPTIMIZATION

4.1 Introduction

In chapter 2 the theory of modelling biomolecular dynamics as a Markov process using a Markov state models (MSMs) was laid out. Central to the method is defining a series of basis sets, $s^i,\ i = 1, ..., n$, which map atomic coordinates to n discrete microstates. These basis sets allow the molecular dynamics trajectories to be represented as an approximate one-dimensional Markov chain from which the MSM can be estimated. The choice of basis set are crucial for accurately capturing the dynamics of system and must therefore be chosen with care [92]. In section 2.3 the process of creating basis sets was described as using four (although many more are possible) modelling choices, or *hyperparameters*, collectively denoted by $\mathbf{x} = (\chi, \tau, m, n)$. Even within reasonable ranges of these hyperparameters, trying each distinct value of $\mathbf{x}$ is computationally intractable. A systematic method is needed to choose appropriate hyperparameters which is reproducible, makes maximum use of the available information, and with the least amount of computational effort. This chapter addresses this need by applying ideas and techniques from the machine learning literature, Bayesian optimisation and response surface methods, and applying them to the test system alanine dipeptide.

Choosing hyperparameters can be thought of as an optimisation problem [132, 300], where the task is to find vectors from hyperparameter configuration space, $X = \chi \times \tau \times m \times n$ which maximise an objective function, f:

$$f : X \rightarrow \text{VAMP-2} \tag{4.1}$$

Within the MM community, efficiently finding the optimum set has recently gained attention [130], however nine out of ten recent studies from a non-random sample[1] performed no hyperparameter optimisation at all.

[1] 10 papers published in 2020 which cite PyEMMA [253] and apply MMs to understand a simulated data set.

Symbol	Definition
$\mathbf{x}$	vector of MSM hyperparameters
χ	MSM hyperparameter: a protein/peptide feature e.g., backbone torsion
τ	MSM hyperparameter: the TICA lag time
m	MSM hyperparameter: the number of TICA components retained
n	MSM hyperparameter: the number cluster centers.
y	the response of an MSM, y = VAMP-2
$\Psi_i(\mathbf{z})$	the right eigenfunction of an MSM
δ_i	the discretisation error of the MSM eigenfunction Ψ_i
$f(\mathbf{x})$	response surface function, a statistical estimation the objective function which is easy to optimise. In the context of Bayesia optimisation called the surrogate function.
$\mathscr{D}_N$	a hyperparameter trial data set. A set of N hyperparameter/response pairs: $(\mathbf{x}_i, y_i)$. Used to estimate the response surface.
τ_{M}	the MSM lag time.
$\mu(\mathbf{x})$	the mean function of a Gaussian Process.
$k(\mathbf{x},\mathbf{x}')$	the covariance kernel of a Gaussian Process. Defines the covariance between the response at $\mathbf{x}$ and $\mathbf{x}'$.
$\mathbf{K}$	the covariance matrix of the Gaussian Process. $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.
θ	the collection of kernel hyperparameters.
η	Kernel hyperparameter: determines the scale of fluctuations of the response.
σ_n	Kernel hyperparameter: determines the noise associated with a single trial.
l_i	Kernel hyperparameter: the characteristic length-scale of the Gaussian process along the ith predictor (MSM hyper-hyperparameter).
R_i	Kernel hyperparameter: the relevance of the ith predictor, $R = \frac{1}{l}$.
$\sigma(\mathbf{x})$	the width of the Gaussian process at the point $\mathbf{x}$. These values are the diagonal elements of $\mathbf{K}$. Depending on context this may or may not include the contribution from σ_n.
$\alpha_{EI}(\mathbf{x})$	the expected improvement acquisition function used to determine the next hyperparameter trial.
(ϕ, ψ)	peptide feature: the backbone torsional angles of an amino acid.
(ϕ, ψ, χ)	peptide feature: the backbone and residue torsional angles of an amino acid.
$\lvert\mathbf{r}_1 - \mathbf{r}_2\rvert$	peptide feature: all heavy atom interatomic distances.
(x, y, z)	peptide feature: Cartesian coordinates
$C_\alpha - C_\alpha$	peptide feature: the alpha carbon contact distances
$X - X$	peptide feature: the heavy atom contact distances

TABLE 4.1: IMPORTANT SYMBOLS USED THROUGHOUT THIS CHAPTER.

Within the larger machine learning community, however, finding the optimum set of hyperparameters for a given model and data set is a common task and has received a lot of attention, as discussed in section 1.6 of the introduction. Much of the focus has been on creating algorithms which automatically optimise hyperparameters rather than a domain expert choosing the values "by hand" [132]. Automated methods of selecting hyperparameters are beneficial for a number of reasons as they [132]:

1. reduce the human and energy resources needed for creating an accurate model,
2. improve the performance of model in general,
3. improve reproducibility and transparency in the model estimation process.

There are two general approaches to hyperparameter optimisation: i) model-free and ii) model-based optimisation [132].

Model-free optimisation techniques include [132]: grid search (or full-factorial design [301], i.e., placing a regular grid over the hyperparameter search space and evaluating each point), random search (i.e., randomly sampling hyperparameters from the search space) and population techniques. The latter include evolutionary algorithms [302], particle swarm optimisation [303, 304] and covariance matrix adaption [305]. When hyperparameter optimisation is performed within the MM community, the former two methods are popular. For example, in reference [92] the authors use random search to determine trends and heuristics for creating MSMs of fast folding proteins, while the authors of reference [257] used grid search over different protein features, various TICA hyperparameters, and number of microstates to optimise a MSM to describe the conformational landscape of the methyltransferase, SETD8. Random search has a number of advantages over grid search. First, when only a small proportion of the hyperparameters are relevant for determining the model score, random search has been shown to be more efficient than than grid search [306]. The reason is that grid search places equal importance on each hyperparameter and effectively wastes the computational budget on combinations of hyperparameters which will score similarly. Second, it is easily adapted to parallel computer architectures and third, increasing the optimisation 'budget' (the number of optimisation steps available) or the size of the search space is easily incorporated into the workflow [132].

Model-based search techniques involve estimating a statistical approximation to the objective function, known as the *response surface* (or surrogate function) and using the response function to choose the next hyperparameter to evaluate [133]. The evaluated hyperparameter is then used to augment the data used to estimate the response surface [133]. The alternating sequence of response surface estimation and hyperparameter evaluation is continued until a satisfactory convergence in the maximum of the response surface is reached [133]. An example response surface for a model with a single hyperparameter is shown in figure 4.1. Evaluating the fictitious model with a hyperparameter value of x leads to a model score of y. The pair (x, y) will be referred to as a *hyperparameter trial*. Repeating trials with different values of x gives a (hyperparameter) *trial data set* $\mathscr{D} = \{(\mathbf{x}_i, y_i)\}$. In the figure the elements of $\mathscr{D}$ are shown as black crosses. The score is a random variable, Y, which

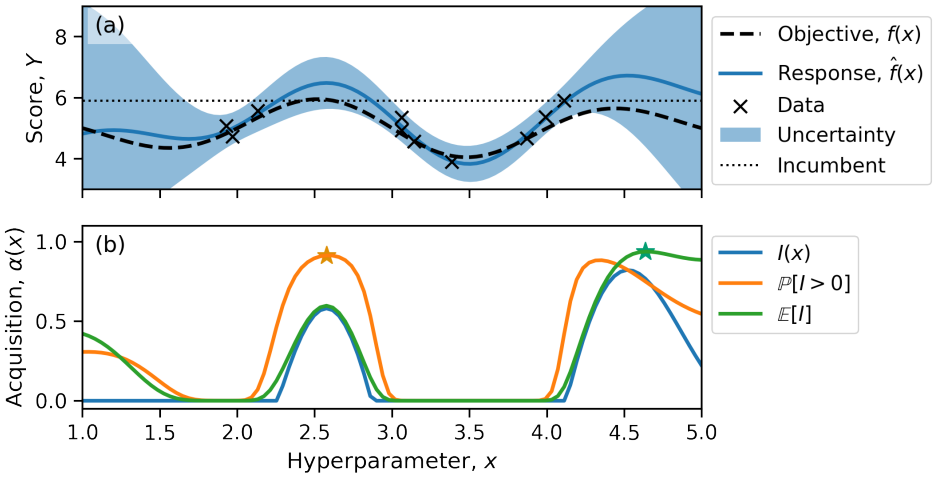


FIGURE 4.1: RESPONSE SURFACE AND ACQUISITION FUNCTIONS. Panel (a) shows the elements of the hyperparameter trial data set, $\mathscr{D}$, as black crosses, the objective function $f(x)$ (black dashed line), and the estimated response surface $\hat{f}(x)$ (solid blue line) with uncertainty (shaded blue area). The dotted line shows the incumbent, $\max\big(f(x)\big)\ x \in \mathscr{D}$. Panel (b) shows a two acquisition functions: the probability of improvement $\mathbb{P}(I > 0)$ (orange line) and the expected improvement $\mathbb{E}[I]$ (green line). For reference the improvement function, $I(x)$ (blue line) is also shown. The maximum of the acquisition functions are denoted with a star.

can be modelled by: $Y \sim f(x) + \epsilon$ where ϵ is an error term. The estimated function, $\hat{f}(x)$, is the response function and is shown as a blue line, while the uncertainty (as captured by ϵ) is shown as a blue shaded area. The uncertainty arises from any random element in evaluating the score (e.g., from cross-validation) or from the model itself.

Bayesian optimisation is a popular model-based technique for optimising hyperparameters [132–136]. The key components of Bayesian optimisation are: i) the *acquisition function*, α, which determines the utility of choosing a particular hyperparameter value, and ii) the response function, which encapsulates all current knowledge of the objective function. The next *candidate* hyperparameter in the optimisation sequence is chosen by maximizing the acquisition function [138]. Acquisition functions trade-off exploration of the search space with exploitation of areas more likely to optimise the objective function. Each does this in their own way, with their own particular strengths and weaknesses, but they come in three main categories i) improvement-based policies, ii) optimistic policies, and iii) information-based policies [138].

Improvement based policies use the improvement function, $I(\mathbf{x})$, (shown in figure 4.1 panel (b) as the blue line), which is defined as the difference between the value of the response surface, $f(\mathbf{x})$, and incumbent, μ^* [138]:

$$I(\mathbf{x}) := \big(f(\mathbf{x}) - \mu^*\big)\mathbb{I}_{f(\mathbf{x})>\mu^*}. \tag{4.2}$$

The incumbent [137] is defined as $\mu^* = \max\big(f(\mathbf{x})\big)$, s.t. $\mathbf{x} \in \mathscr{D}$, and $\mathbb{I}$ is an indicator function which

ensures that $I \geq 0$. In other words, the incumbent is optimum of the response surface but restricted to values of $\mathbf{x}$ which occur in the data used to estimate the response surface. Examples of such policies include probability of improvement [307] $\alpha_{\mathrm{PI}}(\mathbf{x}) = \mathbb{P}[I(\mathbf{x}) > 0]$ (orange line figure 4.1 panel (b)) and the expected improvement [308] $\alpha_{\mathrm{EI}}(\mathbf{x}) = \mathbb{E}[I(\mathbf{x})]$ (green line in figure 4.1 panel (b)). The probability of improvement tends to exploit known regions of the response surface and so can fail to explore sufficiently to find the global optimum [309]. This is because, as can be seen from the definition, it treats all improvements the same no matter how small (although the incumbent can be altered to modify the explore/exploit trade-off) [309]. The expected improvement corrects this by taking into account both the probability of improvement *and* the size of improvement. The difference between the two acquisition functions is demonstrated in figure 4.1. The improvement (blue line, panel (b)) has two peaks, a smaller peak at $x \simeq 2.6$ and a larger peak at $x \simeq 4.6$. The response function at $x \simeq 2.6$ has smaller uncertainty because of the larger number of observations surrounding it. The probability of improvement (orange line, panel (b)) measures the amount of uncertainty above the incumbent line (dotted line, panel (a)) and is greatest here as the majority of the blue shaded area is above the incumbent line. At $x \simeq 4.6$ and beyond, the uncertainty is large but more evenly distributed above and below the incumbent line, hence the probability of improvement peaks and then decreases. However, the expected improvement (green line) is large because the uncertainty in the response extends to high values of y and the expectation is taken from above the incumbent line. The expected improvement is a popular choice [132] and is the default option in some well-known packages such as Spearmint [134] and BayesOpt [140].

Optimistic policies, such as the upper confidence bound [310], choose the candidate to maximize a particular quantile of the uncertainty in the response surface. They have been shown to minimize the 'regret' accumulated over all iterations of the optimisation procedure [310]. The 'regret' being the difference between the true maximum of the objective function and the objective function measured on the candidate hyperparameter trial [311]. More detail on the performance of improvement and optimistic acquisition functions can be found in reference [309]. Information based policies are based on the distribution over the potential hyperparameters, $p(\mathbf{x}|\mathscr{D})$, which describe the probability of optimising the objective function and are induced by the uncertainty in the response surface [138]. Examples include entropy search [312], predictive entropy search [313], and entropy search portfolio [314].

The second component of Bayesian optimisation is the functional form of the response function. Stochastic processes such as Gaussian processes (GPs) or T-student processes (TPs) [149] are common and are implemented in a number of packages [134, 140, 145–148]. Gaussian processes in particular have many useful properties for Bayesian optimisation [132, 137, 300]. First, the improvement and optimistic acquisition functions discusses previously have simple analytic forms [137]. Second, they do not specify a particular form of mean response (unlike for example, general linear models which are linear functions of their predictors [315]). Rather, they specify the structure of the covariance between values of the response through a kernel function $k(\mathbf{x}, \mathbf{x}')$ [149]. This

allows easy fitting of arbitrarily shaped response functions (the response function in figure 4.1 was a Gaussian process model). Third, with recent work in sparse estimation methods they are also able to handle large data sets [316]. Despite their many advantages, GPs do suffer some minor drawbacks and technical hurdles for hyperparameter optimisation. First, they perform poorly with large numbers of categorical hyperparameters [317] compared to tree based response surface models such as tree Parzen estimators, TPEs [135], and random forests, RFs [133, 318]. Second, they come with their own modeling choices which must also be determined [149]. These are the functional form of the covariance kernel, $k(\mathbf{x},\mathbf{x}')$, and transformations of the predictors, or input warping [143].

Estimating the response function of a statistical model is not only beneficial as part of Bayesian optimisation but it also facilitates understanding the effect of hyperparameters on a model, which has lead to important insight for model optimisation [149]. The authors of reference [306] used GPs to demonstrate the important result discussed earlier that random search is more efficient than grid search. They did this by randomly selecting hyperparameters of a deep learning image classifier and for each selection determined the classification accuracy of the model. A Gaussian process was used to model the classification accuracy as a function of the model hyperparameters and the learned parameters of the GP were used to calculate the *hyperparameter relevance*. Hyperparameters with large relevance are important for determining the model score (accuracy in this case). They were able to show that the learning rate (the rate at which the deep learning algorithm updates its parameters in light of new information) was the most relevant hyperparameter. This explained why exploring sets of hyperparameters with the same value of the learning rate, which happens in grid search, differed only slightly in their accuracy. However, they also noted that some hyperparameters were more relevant depending on the nature of the data used to train the model. Similar ideas can be used with other types of surrogate model for the response surface. For example in reference [319] and [320] the authors used random forests to assess the importance of variables for optimising compiler options and machine learning models respectively.

The preceding discussion has highlighted the range of different optimisation techniques available for hyperparameter optimisation: from well established Bayesian optimisation with improvement based policies (which date back to the 1960s) to approaches from in the last 10 years using information theory, such as entropy search. The population technique of covariance matrix adaption, in particular, has performed well in recent benchmarking exercises [321–323]. However, the focus of this chapter will be Bayesian optimisation using improvement based policies, given its long established nature and the fact that no similar techniques have yet been applied to MSMs.

The overall aim of this chapter is develop methods which can help create Markov state models in a more efficient and robust way. That is, given a set of simulation data how can an optimum set of hyperparameters be discovered with as little computational effort as possible? Additionally, how sensitive is this optimum to slight changes in the hyperparameters? To accomplish this aim, the use of Bayesian optimisation and Gaussian processes regression to optimise an MSM from a given set of simulation data will be investigated. Bayesian optimisation is used to optimise machine learning

models, but can it be used optimise Markov state models of biomolecular dynamics? There are a wide range of policies and surrogate functions that have been used for Bayesian optimisation, so to answer this question, empirical tests of promising combinations must be performed. This chapter performs such a test on a popular combination: a policy of optimising the expected improvement using a Gaussian process surrogate function. It would also be beneficial to quantitatively understand which hyperparameters are important for optimising MSMs: if only one or two modelling choices are actually important, this significantly reduces the effort required choose optimal hyperparameters. Gaussian processes regression has also been shown to be useful in this regard and will be used to quantify the relevance of MSM hyperparameters. This also opens up the possibility of performing systematic sensitivity tests, although this will be deferred to chapter 6. Conclusions drawn along the way will be used to suggest practical steps to improve estimating optimal MSMs.

This chapter will investigate the utility of Bayesian optimisation and modelling the MSM response surface, using Gaussian processes regression, of the benchmark system alanine dipeptide (see for example any number of MSM method papers e.g., [91, 168, 324]). This system is well known and the free energy surface is accurately described by just two features: the two backbone dihedral angles [325]. This provides an ideal testing ground for MSM optimisation techniques as at least one optimum hyperparameter is already known (the optimum protein features). This is in contrast to more complex systems such as larger peptides and proteins where the optimum hyperparameters are not in general known, they may involve linear or non-linear combinations of many dihedral angles, for example. This fact, along with the the small hyperparameter space (only two hyperparameters, the protein feature and the number of microstates, are necessary), limits the conclusions of this chapter: more complex free energy surfaces may benefit from different optimisation techniques. However, the purpose of this chapter is to i) practically demonstrate and explain how to use a Gaussian process to model response surfaces, ii) comment on the interpretation of Gaussian processes in understanding the relevance of MSM hyperparameters, and iii) demonstrate and comment on a simple Bayesian optimisation method for optimising hyperparameters. This will lay the ground work necessary for tackling the more complex system of aromatic amine dehydrogenase (AADH) in chapter 6. The chapter is structured as follows: section 4.2 discusses the methods of Gaussian process regression modelling and Bayesian optimisation in detail, section 4.3 presents and discusses the results and section 4.4 discusses conclusions and limitations of this work.

4.2 Methods

4.2.1 Overview

The workflow and methods of this chapter will now be summarised. The data for this chapter is publicly available molecular dynamics (MD) data set used for benchmarking molecular kinetics methods and is described in section 4.2.2. Using this MD data a new *hyperparameter trial data set* was created by estimating Markov state models with different hyperparameters and scoring them

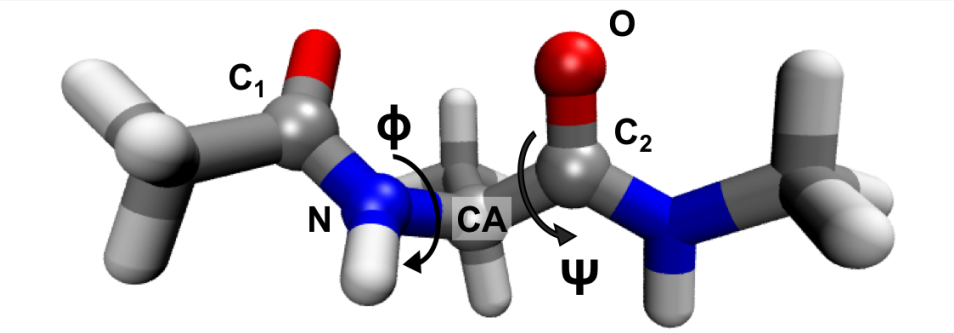


FIGURE 4.2: STRUCTURE OF ALANINE DIPEPTIDE AND DEFINITION OF DIHEDRAL ANGLES. Grey, blue, red and white colors represent carbon, nitrogen, oxygen and hydrogen atoms respectively. The atoms involved in the ϕ, ψ dihedral angles are labeled and highlighted as spheres. The ϕ angle is formed from the intersection of the planes formed by the atoms (C_1, N, CA) and (N, CA, C_2). The ψ angle is formed from the planes formed by the atoms (N, CA, C_2) and (CA, C_2, O).

using the VAMP-2 score. This was performed using both original code and adaptions of open-source packages. A description of these codes and their development is presented in section 4.2.3. The response of an MSM with respect to its hyperparameters was estimated by fitting a Gaussian process (GP) regression model to observations in the hyperparameter trial data set. The theory of GPs is described in section 4.2.4. GPs specify the functional form of the covariance between the dependent variable at different values of the independent variable, this function is called a covariance kernel. The kernel affects how well the regression model fits the data and so models with different kernels were fit and compared using specific goodness-of-fit metrics which are described in section 4.2.5. Parameters of the GP estimated from the hyperparameter trial data set describe how sensitive the VAMP-2 scores are to MSM hyperparameters. A specific measure of this sensitivity called the *relevance* is used in this work and is described in section 4.2.7. Bayesian optimisation (BO) was applied to see if more optimal hyperparameters for an MSM could be found. BO requires an regression model of the response surface and an acquisition function which estimates the expected utility in trying new hyperparameters. The BO algorithm and acquisition function used in this work are described in section 4.2.8.

4.2.2 Molecular dynamics

A molecular dynamics (MD) data set of alanine dipeptide was taken from reference [326]. This data set has been used a benchmark for a number of molecular kinetic methods [98, 324, 326–332]. It consists of $3 \times 250\,\mathrm{ns}$ trajectories sampled from a constant volume, constant temperature ensemble at $T = 300\,\mathrm{K}$ controlled using a Langevin thermostat in explicit (TIP3P [333]) water [326]. The total simulation time is approximately 577 times the longest relaxation timescale of alanine dipeptide (approximately $1.3\,\mathrm{ns}$ [325]) and so all dynamic processes were assumed to be well sampled. The

sampling was performed using the ACEMD [334] package, with the AMBER ff-99SB-ILDN [335] force field and a $2\,\mathrm{fs}$ time-step [326]. Electrostatic forces were computed using the particle-mesh Ewald (PME) [336] summation method every two time-steps with real-space cutoff $9\,\text{Å}$ and grid spacing $9\,\text{Å}$, and all bonds to hydrogen atoms were constrained [326]. The atomic coordinates were saved every $1\,\mathrm{ps}$ and the three trajectories were split into $750 \times 1\,\mathrm{ns}$ smaller trajectories of 1000 frames each.

4.2.3 MSM fitting and scoring

In order to model the response surface of an MSM, a hyperparameter trial data set $\mathscr{D} = \{(\mathbf{x}_i, y_i)\}$ was created. This was created by randomly sampling hyperparameters, $\mathbf{x}$, building an MSM using $\mathbf{x}$, and then measuring the MSM response, y, using the VAMP-2 score.

The VAMP-2 score is a suitable measurement of the quality of the model hyperparameters and can be compared between different models. This fact has formed the basis of a number of studies [92, 130, 246] looking at the effect of different modelling choices on Markov state models (see sections 2.3.2 and 2.3.6 for discussions of these references). Each set of hyperparameters defines a different set of basis states which are used to represent the same underlying MD data. The definition of the basis states directly affects the definition of the eigenvectors of the transition matrix. These eigenvectors are the optimal set of reaction coordinates for representing the slow dynamics present in the MD data [91]. *VAMP-2 measures how well a restricted number of eigenvectors, the slow eigenvectors, describe the underlying dynamics of the simulation data. These eigenvectors are affected by the definition of the basis states and therefor the hyperparameters used to create the basis states*. Maximizing the VAMP-2 score can be thought as minimizing two different types of error (although they are related, see reference [97]). The first is the error arising from describing the dynamics using only the slow dynamical processes. In other words: if we truncate our description to just a handful of eigenvectors how different is this to the true description? If this truncation error is small then the slow eigenvectors provide a satisfactory description, which in turn validates the microstate basis used to estimate these eigenvectors. Maximizing the VAMP-2 score minimizes this error [97]. The second error can be seen by considering the the space spanned by the MSM eigenvectors after scaling by their eigenvalues (kinetic mapping, see section 2.3.2). Distances in this new space, the kinetic distance, define how slowly two conformations interconvert [90]. Using only the slow eigenvectors will reduce the accuracy of this kinetic distance (compared to using all of the eigenvectors). However, maximizing the VAMP-2 score by changing the definition of the basis states, decreases this error and maximizes the proportion of the kinetic variance captured by the slow eigenvectors.

To put the preceding discussion in more salient terms. Consider the same MD data projected onto the same continuous feature, e.g., ϕ dihedral angle, and then onto five (model 1) and 10 (model 2) discrete microstates. In principle the continuous feature could be different as well. There are two slow dynamical processes that need to be described well by the model. Thus, the VAMP-2

score is defined with three eigenvectors (the first one describes the equilibrium distribution and contributes only a constant to the overall score, but is nevertheless included). The VAMP-2 score for model 1 is 2.5, while for model 2 it is 2.9, thus model 2 is preferred. The 10 basis states in model 2 represent a smaller area of the feature space and can capture variations in the eigenvectors more accurately than with just five basis states. The dynamics described by just three eigenvectors from model 2, would be more similar to the true dynamics than the three eigenvectors from model 1.

The hyperparameter search space of alanine dipeptide is shown in table 4.2. It consists of only two hyperparameters, the peptide feature, χ, and the number of cluster centres, n. The structure of alanine dipeptide is shown in figure 4.2 where the two dihedral angles used as features are also labelled. For each value of χ, 100 values of n were randomly sampled and scored. This number was chosen to ensure variation in the response with respect to n was captured. A similar study [92] used between 33 and 100 randomly sampled values per hyperparameter. This meant $\mathscr{D}$ contained $N = 500$ hyperparameter trials.

The response of each trial was measured by building an MSM with a lag time of $\tau_{\mathrm{M}} = 9\,\mathrm{ps}$ and evaluated using VAMP-2 scored with the first $r = 5$ eigenvalues, in line with reference [168]. 20 iterations of 50:50 shuffle-split cross-validation, described in algorithm 1, was used when estimating the VAMP-2 score.

The calculations for this chapter were performed in two stages:

1. **Creation of hyperparameter trial data set**: Markov state models with randomly sampled hyperparameters were estimated and scored. To manage this process the open-source software packages Osprey [148] and PyEMMA [253] were adapted by the author of this thesis. This data was used to estimate the response surface of MSMs of alanine dipeptide.
2. **Response surface estimation and Bayesian optimisation**: The response surface was estimated and Bayesian optimisation was performed also using code developed by the author of this thesis.

These two sets of calculations are described below.

Creation of hyperparameter trial data set

The sampling of hyperparameters and fitting of MSMs was managed by a development version of Osprey (version 1.2.0dev) [148] adapted by the author of this thesis and made available on the code sharing platform GitHub (link to code repository: https://github.com/RobertArbon/osprey). The main changes made were: fixing programming bugs and addition of code to allow compatibility with the package PyEMMA. The fitting and scoring of the MSMs was performed with a development version of PyEMMA (version 2.4) [253] adapted by the author of this thesis to be compatible with Osprey, also available on GitHub (link to code repository: https://github.com/RobertArbon/PyEMMA). The main changes made were to make all the programming classes used in PyEMMA compatible with the application programming interface used by Osprey. Python version 3.5 was used throughout.

```
estimator:
  eval: Pipeline([('cluster',
                KmeansClustering(n_clusters=1,
                                 max_iter=1000,
                                 stride=10)),
                ('msm', MaximumLikelihoodMSM(lag=9,
                                             score_k=5,
                                             score_method='VAMP2'))])
  eval_scope: pyemma

strategy:
  name: random

search_space:
  cluster__n_clusters:
    min: 10
    max: 1000
    type: int

cv:
    name: shufflesplit
    params:
      n_splits: 20
      test_size: 0.5

dataset_loader:
  name: numpy
  params:
    filenames: *.npy

trials:
  uri: sqlite:///osprey-trials.db
  project_name: psi

```

FIGURE 4.3: EXAMPLE OSPREY CONFIGURATION FILE FOR SAMPLING AND SCORING HYPERPARAMETERS. This specifies randomly sampling the number of cluster centers, clustering feature trajectories using k-means, estimating an MSM, scoring using the VAMP-2 score with 50:50 shuffle split cross-validation.

The combination of Osprey and PyEMMA code developed here allows the user to score maximum likelihood Markov state models using the following workflow:

1. Extract features from molecular dynamics trajectories (for example using MDTraj [337]) and store in NumPy arrays [338].
2. In an Osprey configuration file specify:
   a) a trajectory preprocessing (e.g., TICA and clustering) and MSM estimation pipeline using the PyEMMA syntax;
   b) the method for hyperparameter sampling: random sampling, grid search, or Bayesian optimisation;
   c) the hyperparameter search space;

Hyper-parameter	Type	Range	Dim.	Details
Feature, χ	Categorical	(1) (ϕ, ψ)	2	Torsions
		(2) (x, y, z)	30	Heavy atoms only
		(3) ϕ	1	Torsion
		(4) ψ	1	Torsion
		(5) RMSD	1	Heavy atoms only
Cluster centres, n	Integer	10, 11 ... 1000		Clustered using k-means clustering

TABLE 4.2: HYPERPARAMETER SEARCH SPACE OF ALANINE DIPEPTIDE. Prior to feature selection the Cartesian coordinates of the MD trajectories were first aligned to a single, randomly chosen, trajectory frame so that features (2) and (5) did not include spurious rotational or translational motion. The number of dimensions, 'Dim.', refers to the number of individual feature variables created by χ.

   d) the type of cross-validation (e.g., shuffle-split) along with the number of cross-validation iterations.

3. Osprey can then be used to sample hyperparameters, score Markov state models and store the results.

An example Osprey configuration file is shown in figure 4.3. This configuration file loads trajectories of a protein feature stored in NumPy [338] arrays (with `npy` file extension). K-means clustering is performed using the PyEMMA `KmeansClustering` class. A maximum likelihood Markov state model is estimated using the PyEMMA class `MaximumLikelihoodMSM`, with a Markov lag time of `lag`=9 frames. Model scoring is performed using the VAMP-2 score (`score_method='VAMP2'`) with 5 eigenvalues (`score_k`=5). Hyperparameters are selected at random (`name: random`) by selecting the number of cluster centers (`cluster__n_clusters`) from the interval $[10, 1000]$. Cross-validation (`cv`) using the shuffle split algorithm (`name: shufflesplit`) with 20 iterations (`n_splits: 20`) and a test-train data split of 50 % (`test_size: 0.5`) is used to score the models. Information on how to run Osprey can be found in the accompanying documentation, see reference [148].

Bayesian optimisation and analysis

Although the Osprey code can perform Bayesian optimisation, the Bayesian optimisation and estimation of MSM response surfaces were performed separately using code developed the by author of this thesis. All code for this chapter can be found on Github (link to code repository: https://github.com/RobertArbon/alanine_dipeptide). This was written in Python 3.7 using the packages PyEMMA (version 2.5) [253], MDTraj (version 1.9) [337], NumPy (version 1.19) [338], Pandas (version 0.23) [339], Matplotlib (version 3.3) [340], Seaborn (version 0.10) [341] and the Jupyter Project (version 4.6) [342].

4.2.4 Gaussian process regression

A Gaussian process is a distribution over functions [149]. In other words, drawing a sample from a GP returns a mapping from $\mathbf{x}$, an, in general, multidimensional input variable, to a continuous output variable $f(\mathbf{x})$. Considering this function at a set of discrete points, $\mathbf{x}_1, \mathbf{x}_1, \ldots, \mathbf{x}_N$, will give a set of random variables, which together form a multivariate normal distribution:

$$\begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_N) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(\mathbf{x}_1) \\ \vdots \\ \mu(\mathbf{x}_N) \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}_1,\mathbf{x}_1) & \cdots & k(\mathbf{x}_1,\mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_N,\mathbf{x}_1) & \cdots & k(\mathbf{x}_N,\mathbf{x}_N) \end{bmatrix} \right) \tag{4.3}$$

Equation 4.3 is a realisation of a Gaussian process at a set of discrete input points. At each input value, $\mathbf{x}$, there is an associated Gaussian random variable, $f(\mathbf{x})$, with a mean $\mu(\mathbf{x})$, a variance $k(\mathbf{x},\mathbf{x})$, and a covariance between f at $\mathbf{x}$ and $\mathbf{x}'$ given by $k(\mathbf{x},\mathbf{x}')$. The function k is called the covariance function or *kernel*. Equation 4.3 can be written succinctly as $\mathbf{f} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$ [149]. A salient example[2] is that of daily atmospheric carbon dioxide levels at the Mauna Loa observatory. $\mathbf{x} = x$ is a variable representing time (measurements are daily) while $f(x)$ are the daily carbon dioxide levels. The levels rise and fall over the course of the year meaning daily measurements are correlated and this is captured in the function $k(x, x')$. A GP can be subject to some random additive noise, ϵ, and this is written $y = f(\mathbf{x}) + \epsilon$, or succinctly as $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \sigma_n^2 \mathbf{I})$, where $\mathbf{I}$ is the identity matrix and σ_n^2 is the variance of the noise [149] (the mean of the noise is assumed to be zero). Continuing the example, ϵ, would represent random errors in the measurement of carbon dioxide levels.

In order to make the link between the theoretical construct of a GP (equation 4.3) and modelling data, two ingredients are needed: a kernel function $k(\mathbf{x},\mathbf{x}')$ and method of incorporating observations, $(\mathbf{x}_i, y_i)$, $i = 1 - N$, where N is the number of observations. To center this discussion consider the Gaussian kernel [149]:

$$k(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{\left|\mathbf{x} - \mathbf{x}'\right|^2}{l^2} \right) \tag{4.4}$$

This states that values of the GP will be highly correlated between values of $\mathbf{x}$ which are close together and less correlated when they are separated, relative to the value of l, the characteristic length-scale of the GP [149]. In other words, it determines how rapidly y changes for changes in $\mathbf{x}$. Continuing the previous example, if $l = 10$ days, then carbon dioxide levels measured less than 10 days apart would be, on average, highly similar. In contract, two measurements separated by a month would, on average, show little or no correlation with each other.

The process of fitting a GP model, $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \sigma_n^2 \mathbf{I})$, using a Gaussian kernel in the definition of $\mathbf{f}$ (equation 4.4), can be thought of as:

[2] this example is used in reference [149] and is the 'hello world' example of GP regression modelling, frequently used in other texts and probabilistic programming tutorials e.g., reference [343]

1. choosing values of $\mu(\mathbf{x}_i)$,
2. a value of l, which in turn completely defines $k(\mathbf{x},\mathbf{x}')$ (this is a kernel hyperparameter), and
3. and a value the variance of σ_n,

which are consistent with the training data $\mathscr{D} = \{(\mathbf{x}_i, y_i)\} = (\mathbf{y},\mathbf{X})$ and all prior knowledge of the system. In order to do this, Bayes' rule is used [149]:

$$\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X}) = \frac{\mathbb{P}(\mathbf{y}|\mathbf{f},\mathbf{X})\mathbb{P}(\mathbf{f})}{\mathbb{P}(\mathbf{y}|\mathbf{x})}. \tag{4.5}$$

The posterior distribution $\mathbb{P}(\mathbf{f}|\mathbf{y},\mathbf{X})$ is the distribution over all possible GPs, $\mathbf{f}$, which are consistent *with the training data* $(\mathbf{y},\mathbf{X})$. In other words, draws from this distribution will now (hopefully) resemble the data. The posterior takes into account the training data through the likelihood function $\mathbb{P}(\mathbf{y}|\mathbf{f},\mathbf{X}) = \mathscr{N}(\mathbf{f},\sigma_n^2\mathbf{I})$. This is the probability of observing the outputs, *given the inputs and a specific GP*, $\mathbf{X}$ and $\mathbf{f}$. The term $\mathbb{P}(\mathbf{f})$ is a probability distribution over all possible functions $\mathbf{f}$, it incorporates all previous knowledge of the system being studied and is known as the prior [252]. In practice this amounts to specifying a value of $\mu(\mathbf{x})$ and in the current example, a distribution of different values of l. If nothing is known about the system, a value of $\mu(\mathbf{x}) = 0$ and a wide distribution of values over l would be appropriate. In effect this would mean both highly correlated and weakly correlated GPs are both *a priori*, likely [252]. The term $\mathbb{P}(\mathbf{y}|\mathbf{x})$ a factor for normalizing the posterior distribution.

The posterior distribution, like the likelihood function, is also a GP and can be written as follows. Let $\bar{f}_*$ be the mean of the posterior GP at some arbitrary point $\mathbf{x}_*$ (which may or may not be in the training data), and let $\mathbb{V}[f]$ be the covariance between value of the GP at this point and all other points in the training data, then [149]:

$$\bar{f}_* = \mathbf{k}_*^\top \left(\mathbf{K} + \sigma_n^2\mathbf{I}\right)^{-1}\mathbf{y} \tag{4.6}$$

$$\mathbb{V}[f_*] = k(\mathbf{x}_*\mathbf{x}_*) - \mathbf{k}_*^\top \left(\mathbf{K} + \sigma_n^2\mathbf{I}\right)^{-1}\mathbf{k}_*. \tag{4.7}$$

Here $\mathbf{k}_*$ is a vector of covariances between $\mathbf{x}_*$ and the training observations, $\mathbf{X}$. As the point $\mathbf{x}_*$ is arbitrary, these equations define the posterior GP over the entire domain.

Equations 4.6 and 4.7 only determine how the posterior GP should be defined in terms of the training data and the hyperparameters[3] of the kernel function, e.g., l in equation 4.4. How can an appropriate value of l be estimated? General kernel hyperparameters (kernels can involve many more parameters) can be estimated from the data using Bayesian estimation or by maximizing the *log marginal likelihood* [149]:

$$\log\left(\mathbb{P}(\mathbf{y}|\mathbf{X})\right) = -\frac{1}{2}\mathbf{y}^\top\left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y} - \frac{1}{2}\log\left|\mathbf{K} + \sigma^2\mathbf{I}\right| - \frac{N}{2}\log 2\pi. \tag{4.8}$$

The values of the kernel hyperparameters estimated by maximizing the log marginal likelihood, equation 4.8, are known as maximum *a posteriori* (MAP) estimates [149]. These are single point estimates of the kernel hyperparameters. When estimates of variability of kernel hyperparameters are

[3] not to be confused with the hyperparameters of the MSM which are the predictors of the GP.

required, Bayesian estimation can be used [252]. In this case Markov chain Monte Carlo (MCMC) is used to sample the posterior distribution of l. The same considerations for Bayesian MSMs in section 2.3.4.3 apply in this case.

The process of fitting a GP to data can be summarised as follows:

1. collect training data $\mathscr{D}$,
2. specify a prior mean function,
3. specify a functional form of covariance kernel $k(\mathbf{x},\mathbf{x}')$, e.g., equation 4.4,
4. specify priors over the kernel hyperparameters,
5. fit the GP by maximizing the log marginal likelihood, equation 4.8 or using Bayesian estimation.

4.2.5 Evaluating model fit

There is considerable flexibility when using GPs to model data. First, there is a wide variety of kernels that can be used, see for example reference [344] which presents a 'cook-book' for constructing complex kernels from other, simpler, kernels. The input variables may also be transformed, a process known as input warping [143], e.g., a logarithmic warping would be to make the replacement $x \rightarrow \log(x)$.

A method for comparing GPs constructed with different kernels and input warpings is needed. For models fit by maximizing the log marginal likelihood the predictive value of the GP can be measured through the standardized mean square error (SMSE) and the mean standardized log loss (MSLL) [149]. These play the same rôle as, for example, the R^2 or deviance play in generalized linear models [315]. The SMSE is defined by [149]:

$$\text{SMSE} = \left(\frac{1}{N}\right)\sum_{i=1}^{N}\frac{\left(f(x)-y_i\right)^2}{\sigma^2_{obs}}, \tag{4.9}$$

and the mean standardized log loss (MSLL) by [149]:

$$\text{MSLL} = \left(\frac{1}{N}\right)\sum_{i=1}^{N}\left[\left(\frac{1}{2}\log\left(2\pi\sigma_i^2\right)+\frac{\left(f(x)-y_i\right)^2}{2\sigma_i^2}\right)-\left(\frac{1}{2}\log\left(2\pi\sigma^2_{obs}\right)+\frac{\left(\bar{y}-y_i\right)^2}{2\sigma^2_{obs}}\right)\right]. \tag{4.10}$$

Here σ^2_{obs} is the observed variance of the training data response, $\mathbb{E}\left[(y_i-\bar{y})^2\right]$, and σ_i^2 refers to the GP predicted variance at the observation i including the noise term (i.e., $\mathbb{V}\left[f_*\right]+\sigma_n^2$, from equation **??**). The 'standardization' in each case defines a baseline model with which to compare the fitted GP model. This baseline, or null, model predicts the mean and variance of the observed y_i at each value of $\mathbf{x}_i$ i.e., [149]:

$$\bar{f}_*^{\text{Null}} = \mathbb{E}\left[y_i\right] \tag{4.11}$$

$$\mathbb{V}^{\text{Null}}\left[f_*\right] = \mathbb{V}\left[y_i\right]. \tag{4.12}$$

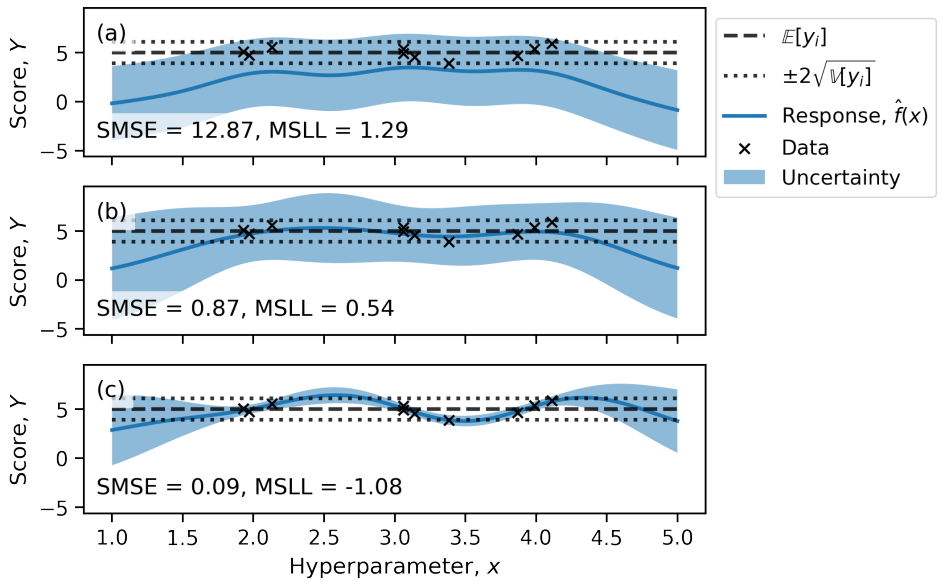


FIGURE 4.4: MEAN STANDARDIZED LOG-LOSS (MSLL) AND STANDARDIZED MEAN SQUARE ERROR (SMSE). Panels (a) - (c) show three different GPs fit to the same data after 3, 6, 9 steps, respectively, of marginal likelihood maximization. The MSLL and SMSE is for each GP is labelled. The blue line and shaded blue area are the mean and uncertainty of the GP, the black crosses are the observations, and the dashed and dotted lines are the mean and uncertainty of the null model (i.e., mean and uncertainty of y_i).

In other words, the SMSE and MSLL both answer the question: how much better is the GP model at making predictions than just fitting a single Gaussian distribution to the response? To gain intuition of the SMSE and MSLL consider figure 4.4. Panels (a), (b) and (c) show a GP fit to the same data as figure 4.1 after 3, 6, and 9 iterations of marginal likelihood maximization. This would not be done in practice but serves as an example of three models which differ in how well they fit the data. In practice these differences would arise from different kernels etc. The null model is denoted by the dashed and dotted lines (mean and uncertainty respectively), while the mean and the uncertainty in the fitted GP are denoted by the blue line and shaded blue area respectively. In panel (a) the fitted GP is clearly worse than predicting the mean of the observations, so $\mathrm{SMSE} > 1$ and $\mathrm{MSLL} > 0$. In panel (b) the GP fits the data well so $\mathrm{SMSE} < 1$ but the variance of the GP is still much larger than the variance of the observations so $\mathrm{MSLL} > 0$. In panel (c) the GP fits the observations almost exactly so $\mathrm{SMSE} \simeq 0$, and the variance of the GP at each observed $\mathbf{x}_i$ is smaller than variance of the observations so $\mathrm{MSLL} < 0$.

In this thesis, the MSLL and SMSE were estimated using K-fold cross-validation [150] to avoid choosing a kernel or input warping which may over-fit to a particular data set. The K-fold cross-validation procedure is as follows [150]: first split the hyperparameter trial data set, $\mathscr{D}$, into a K equally sized, disjoint, sets or 'folds'. Second, fit the GP using $K-1$ folds and then calculate the SMSE and MSLL on the remaining 1 fold. Third, repeat this process K times, training the GP

and calculating the SMSE and MSLL with a different held out fold each time. The cross-validated MSLL and SMSE is the weighted mean of the SMSE and MSLL on each fold.

4.2.6 Response surface modelling

The VAMP-2 response of the MSM with respect to its hyperparameters, $\mathbf{x} = (\chi, n)$, was modelled as a GP with additive noise. A variety of combinations of input warping (e.g., log-transformation of $\mathbf{x}$) and covariance kernels, $k(\mathbf{x}, \mathbf{x}')$, were tried and the best combinations for each response surface determined using the cross-validated SMSE and MSLL.

To reduce the computational effort required to fit each GP model, a sparse approximation to the full covariance matrix of the GP, called the fully independent training conditional (FITC) was used [316]. In this approximation a number of observations must be designated as 'inducing points'. Larger numbers of inducing points increases the accuracy of the approximation at the expense of increased computational effort. The number of inducing points was set to $10\,\%$ of the total number of observations and their location determined by k-means clustering as suggested in the probabilistic programming package PyMC3 (version 3.5) [343]. This fraction was chosen by fitting a single GP model with the number of inducing points ranging from $10\,\%$ to $100\,\%$ of the total observations. The number of inducing points was not found to change the posterior distribution of any kernel hyperparameters significantly and so the smallest value was used.

As described in the previous section, in order to fit a GP model, a number of modelling choices need to be specified, these are:

1. the prior of the mean function, $\mu(\mathbf{x})$;
2. the kernel function, $k(\mathbf{x}, \mathbf{x}')$;
3. the prior distributions of kernel hyperparameters;
4. the warpings of the predictors.

The kernel function and input warping will be chosen by fitting models and selecting the combination which best fits the data. The mean function and prior distributions over kernel hyperparameters will be set based on other consideration.

Mean function and kernel function

The prior of mean function was set to zero everywhere: $\mu(\mathbf{x}) = \mathbf{0}$, in practice this does not have much impact on the final model [137].

The kernel functions considered were restricted to stationary kernels, i.e., those where $k(\mathbf{x}, \mathbf{x}') = k(|\mathbf{x} - \mathbf{x}'|)$. Stationary kernels are advantageous because they admit a useful interpretation of the kernel hyperparameters which will be described in section 4.3.2. Mathematically they mean that the correlations between values of y do not depend on the absolute values of $\mathbf{x}$ only on the distance between $\mathbf{x}$ and $\mathbf{x}'$. The form of kernel used in this work is the same as the one used by authors of

reference [306] in their work on the relevance of hyperparameters discussed in the introduction:

$$k\left(\left|\mathbf{x}-\mathbf{x}'\right|;\theta\right)=\eta^2\prod_i k_M\left(\left|x_i-x_i'\right|;\nu,l_i\right)+\sigma_n^2\delta_{\mathbf{x},\mathbf{x}'}. \tag{4.13}$$

The η terms controls the total variation in the response function, the larger the value of η the more the response is able vary over the whole predictor space [149]. The σ_n^2 term is the noise [149] term which allows the GP to account for variation in the response due to measurement error, which in this case amounts to variation due to the cross validation procedure. The index, i, runs over each component of $\mathbf{x}$ so that x_i refers to a single hyperparameter, e.g., the number of microstates, n, or the peptide feature, χ. The total kernel is the product of kernels over each hyperparameter. The M in k_M denotes a Matérn kernel parameterized by ν - this will be discussed below. l_i is the characteristic length-scale of k_M. η, σ_n and the l_i's are the kernel hyperparameters (collectively denoted by θ) and estimated from the data. The multiplicative form of this kernel means that VAMP-2 responses are correlated only when the values of the predictors are simultaneously similar, where the similarity is set by the value of l_i.

The kernels over the individual predictors were kernels in the Matérn class with values of $\nu = 1/2,\ 3/2,\ 5/2,\ \infty$. These are alternatively known as an exponential, Matérn 3-2, Matérn 5-2 and Gaussian (Radial Basis Function, RBF) kernels. These kernels were chosen based on their common usage [138] and range from 'rough' exponential kernel, where correlation in the response drops off rapidly with changes in the predictors, to smooth processes with the Gaussian kernel. They are defined as follows [149]:

$$k_{\text{Exp}}\left(r;1/2\right)=\exp(-r) \tag{4.14}$$

$$k_{\text{M3-2}}\left(r;3/2\right)=\exp(-\sqrt{3}\,r)(1+\sqrt{3}\,r) \tag{4.15}$$

$$k_{\text{M5-2}}\left(r;5/2\right)=\exp(-\sqrt{5}\,r)\left(1+\sqrt{5}\,r+\frac{5}{3}r^2\right) \tag{4.16}$$

$$k_{\text{RBF}}\left(r;\infty\right)=\exp\left(-\frac{1}{2}r^2\right) \tag{4.17}$$

where $r=\frac{|x-x'|}{l}$. See chapter 5 of reference [149] for a full description of the Matérn kernels and their properties.

4.2.6.1 Prior distributions of kernel hyperparameters

Prior distributions of kernel hyperparameters affect the range of possible values which can be estimated from the data. They serve the purpose of ensuring that the learned hyperparameters fit with prior expectations about their true values [252].

The GP hyperparameters estimated from the data are the mean function, $\mu(\mathbf{x})$ and the kernel hyperparameters, $\theta=(\eta,\sigma_n,l_1,l_2,...)$. These were estimated differently depending on the application. When the GP was used for visualisation (e.g., figure 4.9) or for Bayesian optimisation (section 4.3.3) the hyperparameters were estimated by maximizing the log marginal likelihood. When error

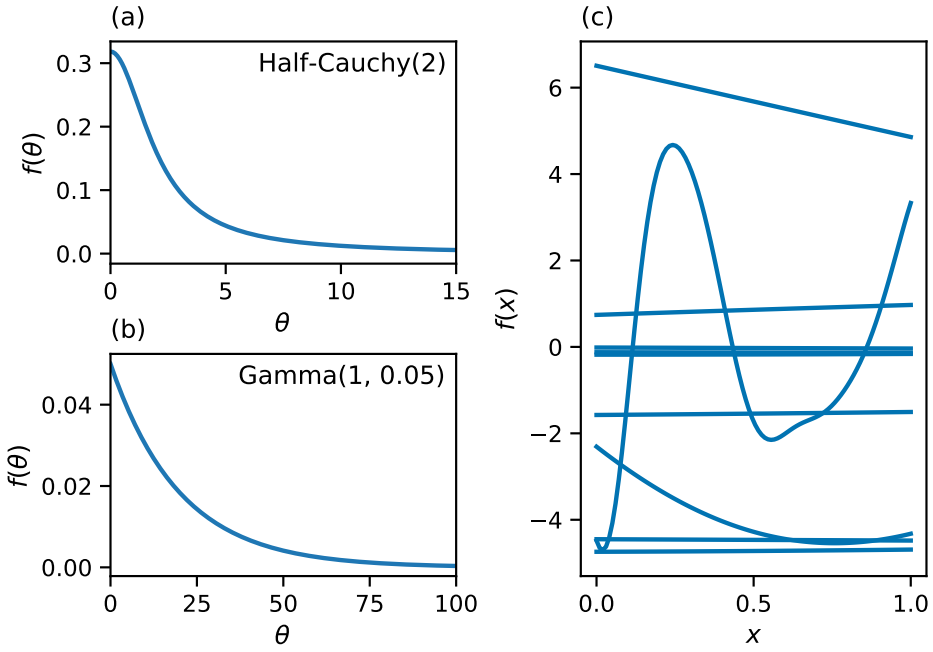


FIGURE 4.5: KERNEL HYPERPARAMETER PRIORS AND REPRESENTATIVE GPS. Panel (a) shows the prior used for the variance parameters, η, σ_n: the half-Cauchy with $\beta = 2$. Panel (b) shows the prior used for the length-scale parameters l_i: the Gamma distribution with $\alpha = 1, \beta = 0.05$. Panel (c) shows 10 draws from a Gaussian process with an RBF kernel (equation 4.17) with values of l drawn from the distribution in panel (b) and values of η drawn from distribution in panel (a).

estimates of the GP hyperparameters were needed for discussing the relevance (section 4.3.2) Bayesian estimation was used.

The prior distributions for the variance terms, η and σ_n, were $\text{half}-\text{Cauchy}(\beta = 2)$ and the priors for the length-scale parameters, l_i, were $\text{Gamma}(\alpha = 1, \beta = 0.05)$. These distributions are shown in figure 4.5 panels (a) and (b) respectively. To get a sense of the effect of these priors on a GP, 10 draws from a 1D Gaussian process with an RBF kernel are shown in 4.5 panel (c). The values of the η and l hyperparameters in this GP are drawn the distributions in panels (a) and (b) respectively.

The rôle of weakly informative priors is to exclude unrealistic or disallowed values of the parameters without imposing strong prior beliefs on the true values [252]. The half-Cauchy distribution was used for η and σ_n based on its recommended use in other settings [345]. It was only necessary for the scale of this distribution to give significant density in the range $0-5$ as the VAMP-2 score will lie in the range $[1,5]$ for alanine dipeptide thus limiting the possible values of η and σ_n. The prior for l was justified on the basis that, after scaling the predictors to lie in $[0,1]$, values of $l \gg 1$ imply a flat response, meaning significant density for $l \gg 1$ is not necessary.

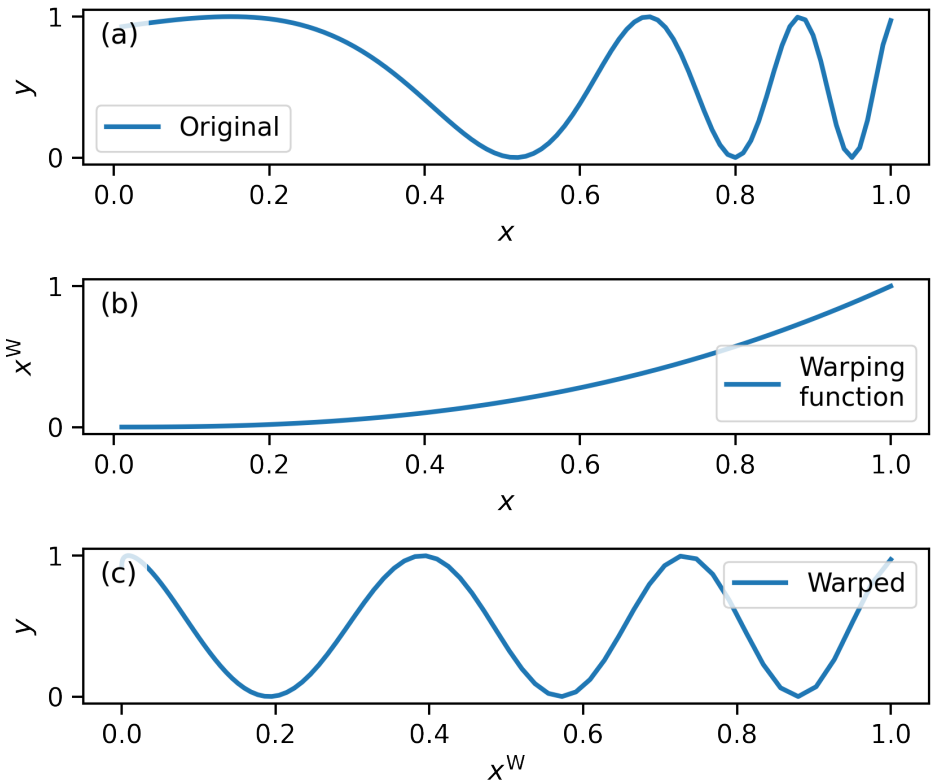


FIGURE 4.6: INPUT WARPING. Panel (a) shows a non-stationary function y as a function of a predictor x. The characteristic length-scale decreases as x increases. Panel (b) shows a warping function which transforms x to x^{W}. Applying this warping results in the new function in panel (c) which is approximately stationary. This figure is an adaption of figure 1 in reference [143].

4.2.6.2 Preprocessing

The values of the MSM hyperparameters, the inputs or independent variables of the GP, were preprocessed in three ways: scaling, input warping, and coding (for categorical predictors only).

All inputs were scaled to lie in the range $[0,1]$. This was to make the kernel length-scale parameters, l, comparable across different hyperparameters. This effectively makes the units of l the same for each input.

Input warping is used to mitigate the problems of modelling non-stationary functions using stationary GP kernels [143]. Assuming stationarity i.e., that the GP characteristic length-scale does not vary with the input $\mathbf{x}$, simplifies both estimation and interpretation of the GP [143]. Thus, warping predictors to make the stationarity assumption more plausible is important, especially when it comes to discussing hyperparameter relevance in section 4.2.7. Coding is used to transform non-numerical predictors (i.e., the protein feature) into numerical variables.

To aide understanding of input warping, a dramatic example is shown in figure 4.6 (which is an adaption of figure 1 in reference [143]). Panel (a) shows a non-stationary function - the characteristic

χ	n		χ_1	χ_2	χ_3	χ_4	χ_5	n^s
(ϕ, ψ)	10	$\longrightarrow$	1	0	0	0	0	0.00
(x, y, z)	500		0	1	0	0	0	0.49
ψ	1000		0	0	0	1	0	1.00

FIGURE 4.7: EXAMPLE PREDICTOR TRANSFORMATION. The left hand table shows the raw values of the hyperparameters, χ and n, as a data matrix. The right hand table shows same values after dummy coding $\chi \rightarrow \chi_1, \chi_2, \ldots$ and scaling $n \rightarrow n^s$ to lie in the range $[0, 1]$. No warping was applied. The dummy coding and scaling is performed with reference to the hyperparameter search space in table 4.2.

length-scale decreases with increasing x. In other words, the kernel function, $k(x, x')$ could not be written as a function of $|x - x'|$. Panel (b) shows a warping function which transforms x to x^w. Panel (c) applies this warping and the resultant function is more plausibly stationary. With respect to this work, two input warpings, $T(x)$, were considered: the identity $I(x)$, and a logarithmic transformation, $\log(x)$.

The categorical predictor, χ, was dummy coded [346] to give a five dimensional vector of 1s and 0s. A example of the transformations of two hyperparameters, χ and n, in preparation for modelling with a GP are shown in figure 4.7.

In order to select the best combination of kernel functions and predictor warping each possible combination was used to estimate a GP which was then evaluated using the cross-validated MSLL and SMSE. So, for the response surface of alanine dipeptide eight different models were estimated: four different types of kernel (equations 4.14 - 4.17) were used in equation 4.13 and two predictor warpings for n. For each model both metrics were calculated using 10-fold cross validation. Any model with MSLL> 0 or SMSE> 1 was discarded. The remaining models were ranked separately according to MSLL and SMSE (R_{MSLL}, R_{SMSE}) and the ranks combined according to $\sqrt{R^2_{\text{MSLL}} + R^2_{\text{SMSE}}}$. This ranking method was used to ensure a balance between the two selection metrics compared to, say, the mean of the two ranks.

All GPR modelling was performed with the Python package PyMC3 (verion 3.5) [343] with some visualisation performed using package GPy (version 1.5) [347].

4.2.7 Hyperparameter relevance

The sensitivity of the outcome y to changes in the predictors is measured by a function of its characteristic length scale called the relevance. The characteristic length-scales in equation 4.13, l, each correspond to a different predictor, or level of categorical predictor. They determine the covariance of the response between points with different values of that predictor. For example, for an exponential kernel with $l = 1$ then inputs separated by $|x - x'| = 1$ will on average have a covariance of $\exp^{-0.1} \simeq 0.9$. This means for large values of l the response with respect to changes in x will be flat, or in other words, x is irrelevant to determining the response. This prompts the definition

of *relevance*, $R = \frac{1}{l}$ [306, 348]: when R is large, the small changes in x result in larger changes in the response, meaning it is relevant to determining the response. Hereafter the kernel functions (equations 4.14 - 4.16) will be parameterized interchangeably with R and l where convenient.

The relevance of the MSM hyperparameters is important for understanding and visualising the response surface and so to calculate the uncertainty in R a fully Bayesian approach was used. After model selection using the maximum marginal likelihood models described in section 4.2.6, the GP model hyperparameters were re-estimated by sampling the posterior distribution using Markov Chain Monte Carlo. A No U-Turn sampling algorithm, using two independent chains with 500 tuning steps and 1000 sampling steps. Convergence was checked using the R-hat statistic [254].

4.2.8 Bayesian optimization

The response surface of an MSM can be optimised using Bayesian optimisation. As discussed in the introduction to this chapter Bayesian optimisation requires two ingredients: i) a response function which models the response of the MSM to its hyperparameters, and ii) an acquisition function. The response function was chosen using the methods outlined in the previous section.

Bayesian optimisation and the acquisition function, α, can be understood by considering two values of the predictor, $\mathbf{x}_1$ and $\mathbf{x}_2$. The goal of Bayesian optimisation is to maximize the function $f(\mathbf{x})$. The values of $f(\mathbf{x}_1)$ and $f(\mathbf{x}_2)$ are unknown. A choice must be made as to which value of $\mathbf{x}$ to use to evaluate next, given that evaluating $f(\mathbf{x})$ is costly and it is not possible to simply try all possible values of $\mathbf{x}$. If the values of the acquisition function are $\alpha(\mathbf{x}_1) > \alpha(\mathbf{x}_2)$ then $f(\mathbf{x}_1)$ is expected to be greater than $f(\mathbf{x}_2)$. Acquisition functions are functions of the expected value of $f(\mathbf{x})$ and the variance in this estimate. This is precisely the information provided by the response surface modelled as a Gaussian process (although many different response surface models are possible [149]).

The acquisition function used is the expected improvement, $\mathbb{E}[I]$ where the improvement, I, is defined as [138]:

$$I(\mathbf{x}, \mu^*) := (f(\mathbf{x}) - \mu^*)\mathbb{I}_{f(\mathbf{x})>\mu^*}. \tag{4.18}$$

The improvement at $\mathbf{x}$ (the true response at which is unknown) is how much the estimated value $f(\mathbf{x})$ exceeds some threshold. This threshold is taken to be the highest value the function which is currently known, this is the incumbent, μ^*. Taking the expectation of this for a Gaussian process gives [138]:

$$\alpha_{EI}(\mathbf{x}) := \mathbb{E}\left[I(\mathbf{x}, f(\mathbf{x}), \mu^*)\right] \tag{4.19}$$

$$= (\mu(\mathbf{x}) - \mu^*)\Phi\left(\frac{\mu(\mathbf{x}) - \mu^*}{\sigma(\mathbf{x})}\right) + \sigma(\mathbf{x})\phi\left(\frac{\mu(\mathbf{x}) - \mu^*}{\sigma(\mathbf{x})}\right) \tag{4.20}$$

Here Φ, ϕ are the normal cumulative and probability distribution functions respectively, and $\sigma(\mathbf{x})^2$ is the variance of the GP at the point $\mathbf{x}$. The expected improvement answers the question: for each value of the inputs ($\mathbf{x}$), which value would be expected to give the highest average response ($f(\mathbf{x})$), over and above what is currently the best value (the incumbent, μ^*)?

It is possible to take the expectation over both the distribution of f and of the GP hyperparameters θ. This has been suggested and shown to be effective [134]. However, this was not done in this work because of the extra computational cost involved.

The Bayesian optimisation algorithm [138] starts with a hyperparameter trial data set of size N_{seed} which was used to estimate an initial response surface $f(\mathbf{x};\mathscr{D}_{N_{\text{seed}}})$ and the incumbent calculated, $\mu^* = \max\left[f(\mathbf{x})\right]$, $\mathbf{x} \in \mathscr{D}_{N_{\text{seed}}}$. Using the response surface and the incumbent, the candidate hyperparameter $\mathbf{x}_1$ was chosen by finding the maximum of the acquisition function. The maximum was found first setting up a grid of points over the hyperparameter search space, $\mathbf{X}_M$, with M points per hyperparameter. $\alpha(\mathbf{x})$ was estimated for every $\mathbf{x} \in \mathbf{X}_M$ and the next candidate hyperparameter chosen as the value which maximized the acquisition function: $\mathbf{x}_1 = \arg\max_{\mathbf{x}} [\alpha_{\text{EI}}(\mathbf{x})]$. The response, y_1, of the MSM to this candidate was calculated, and the trial $(\mathbf{x}_1, y_1)$ added to the trial data set, which becomes $\mathscr{D}_{N_{\text{seed}}+1}$. This process is repeated for p steps and is summarised in in algorithm 3.

Algorithm 3: Bayesian Optimisation.

Data: Trial data: $\mathscr{D}_N = \{(y_1, \mathbf{x}_1), ..., (y_N, \mathbf{x}_N)\}$
Data: Search space grid: $\mathbf{X}_{\text{M}} = \{(\chi_1, \tau_1, m_1, n_1), ..., (\chi_M, \tau_M, m_M, n_M)\}$
Result: $\mathbf{x}^* = \arg\max_{\mathbf{x}} f(\mathbf{x};\mathscr{D}_{N+p})$

for $i \leftarrow N$ **to** $N + p$ **do**
- estimate GP response $f(\mathbf{x};\mathscr{D}_i)$;
- calculate incumbent: $\mu^* = \arg\max f(\mathbf{x};\mathscr{D}_i)$ s.t. $(y, \mathbf{x}) \in \mathscr{D}_i$;
- estimate acquisition function: $\alpha_{\text{EI}}(\mathbf{x};\mathscr{D}_i)$ $\mathbf{x} \in \mathbf{X}$;
- select candidate: $\mathbf{x}_{i+1} = \arg\max_{\mathbf{x}} \alpha_{\text{EI}}(\mathbf{x};\mathscr{D}_i)$ s.t. $(\mathbf{x} \in \mathbf{X})$ & $(\mathbf{x} \notin \mathscr{D}_i)$;
- query objective function to obtain: y_{i+1};
- augment data: $\mathscr{D}_{i+1} \leftarrow \{\mathscr{D}_i, (y_{i+1}, \mathbf{x}_{i+1})\}$

It was observed during these experiments that the same candidate hyperparameters were being proposed by the algorithm. This was deemed due to the granularity of the grid used in the maximisation of the acquisition function. To ensure that each candidate hyperparameter is unique, the Bayesian optimisation algorithm was modified so that only hyperparameters sets not already in the trial data set were considered as candidates. This is reflected in the conditions on the 'select candidate' step of algorithm 3.

Although not extensively discussed in the theoretical literature, software packages seed the process with randomly selected hyperparameter trials so that the initial response surface contains some information, rather than just a random draw from the prior function distribution. For example, the default in BayesOpt [140] is $N_{\text{seed}} = 10$. Conventional advice for parametric models (e.g., multi-variable linear regression) puts the required number of observations for estimating parameters as at least 15 observations per parameter [349]. An appropriate number was explored in this work.

4.3 Results and discussion

4.3.1 Response surface of alanine dipeptide

4.3.1.1 Hyperparameter trial data set

Alanine dipeptide undergoes three relaxation processes which are resolvable with the lag time of $\tau = 9\,\mathrm{ps}$ used here, with implied timescales of approximately $1300\,\mathrm{ps}$, $66\,\mathrm{ps}$ and $30\,\mathrm{ps}$. These values were estimated using a Markov state model using the (ϕ, ψ) torsion feature and $n = 100$ microstates and are inline with other studies using this data e.g., [98, 330]. As the eigenvalue associate with each relaxation timescale can be at most 1, there is an upper-bound on the VAMP-2 score of 4 (this includes the eigenvalue of exactly one corresponding to the stationary distribution, see chapter 2 section 2.3.6). The maximum values of the VAMP-2 score ($\simeq 3.25$) are below this bound and similar to the values estimated in reference [98].

The response of a Markov state model to the type of protein feature (χ) and number of cluster centers (n) was measured by the cross-validated VAMP-2 score (f) using the first five eigenvalues of the transition matrix. The hyperparameter trial data set consisted of 500 observations of f and (χ, n) and is shown in figure 4.8. The test response ($f^{\mathrm{test}} = f\left(\chi, n; \mathbf{X}^{\mathrm{test}}\right)$, blue points) and the difference between train and test response, ($\Delta f = f^{\mathrm{train}} - f^{\mathrm{test}}$, orange) are both shown as functions of n. The features are ordered according to the mean of the test response. As expected from previous work [325] the (ϕ, ψ) feature has the highest average response but figure 4.8 also shows that the heavy atom (x, y, z) coordinates feature performs just as well.

The difference between the train and test response, the *over-fitting*, reflects the consistency between the eigenvectors estimated on the training data and those implied from the time-lagged covariance and overlap matrices (C and Π in 2.15) estimated on the test data [94]. So a small Δf implies that the picture of the relaxation processes are represented equally well, with the given hyperparameters, in both the training and test data (even if they are both inaccurate). This is likely due to the large volume of data used to train the MSMs.

4.3.1.2 Gaussian process regression

The response surface (figure 4.9) was modelled as a Gaussian process with χ and n as predictors. A Matérn 5-2 kernel and logarithmic input warping of n were chosen using a combination of the MSLL and SMSE model selection criteria. The SMSE and MSLL of the response surface was 0.0007 and -4.2369 respectively, see table 9.1 for the selection metrics of all the models. The choice of logarithmic warping of n is unsurprising given that the response for the (ϕ, ψ) and (x, y, z) features (panels (a) and (c) in figure 4.8) is a clearly non-stationary process: the covariance of the response with respect to changes in n is much lower for $n \leq 100$ than for $n \geq 100$ where the response reaches a plateau. The log transformation smooths the response with respect to n and makes the assumption of a stationarity more plausible.

The categorical inputs (the protein feature, χ) posed no problems for GP model, the response surface fits the observed data well across each feature, both in terms of the mean response and its

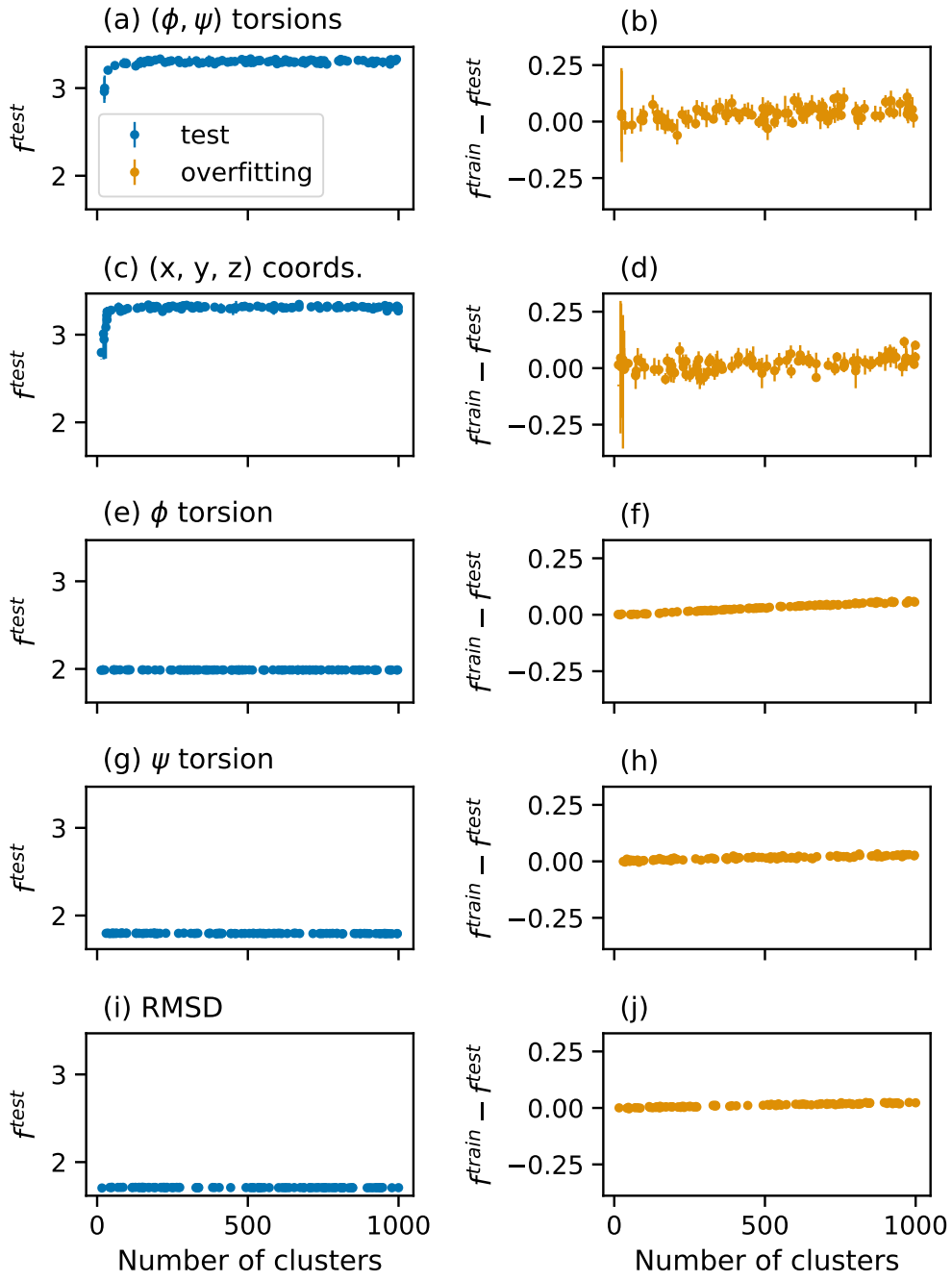


FIGURE 4.8: VAMP-2 SCORES OF THE HYPERPARAMETER TRIALS FOR MSMS OF ALANINE DIPEPTIDE. The test response, $f^{\text{test}} = f(\chi, n; \mathbf{X}^{\text{test}})$ is shown in blue, panels: (a), (c), (e), (g), (i), while the degree of over-fitting, $f^{\text{train}} - f^{\text{test}}$, is shown in orange, panels: (b), (d), (f), (h), (j). Each row represents a different value of the feature, χ, and the horizontal axis represent the number of clusters, n. Each trial was scored with 20 iterations of 50:50 shuffle split cross validation. The error bars represent the 25th and 75th quantiles of the cross-validation folds. The features are ordered according to the mean of the their test scores.

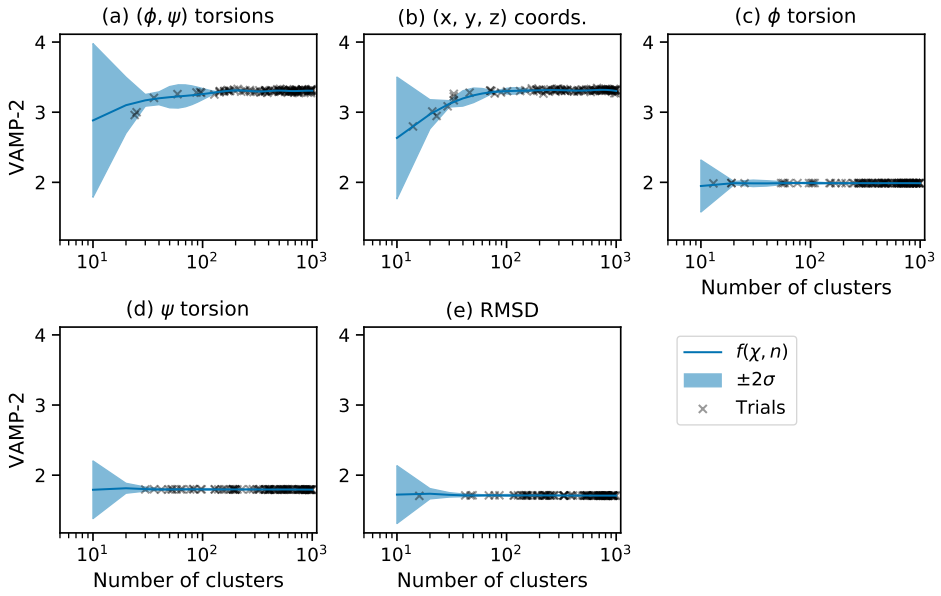


FIGURE 4.9: RESPONSE SURFACE OF ALANINE DIPEPTIDE. The response is shown as a function of the feature, χ (panels (a) - (e)) and number of clusters, n (horizontal axis). The features are ordered according to their average response. A Matérn 5-2 kernel and logarithmic warping of the predictor n was used. The blue line is the mean of the surface, the blue shaded bands represent the uncertainty ($\pm 2\sigma$ excluding the noise term σ_n), and the black crosses are the observed values (the cross validated mean of VAMP-2).

uncertainty.

4.3.1.3 Response surface features

The three aims of this chapter are to demonstrate the ability of GPs to model the response surface of an MSM, use this response surface to understand the sensitivity of MSMs to their hyperparameters, and optimise the response surface with respect to its hyperparameters. The previous section showed the success of the GP model for modelling the alanine dipeptide response surface. Before moving on to the next two aims, there are a number of features of the response surface that are worth discussing. It should be noted that although this simulation data has been used in a variety of molecular kinetics studies [98, 324, 326–332], this is the first explicit estimation of an MSM response surface which means only qualitative comparisons can be to other studies.

First, there is a decrease in response as $n \to 10$ for the (ϕ, ψ) and (x, y, z) coordinate features. Qualitatively this decrease with n is in agreement with previous studies on other systems [94, 97] and is due to increasing eigenfunction discretization error, δ [64] as a n decreases. The discretization error measures the difference between the true eigenfunction and the eigenfunction approximated

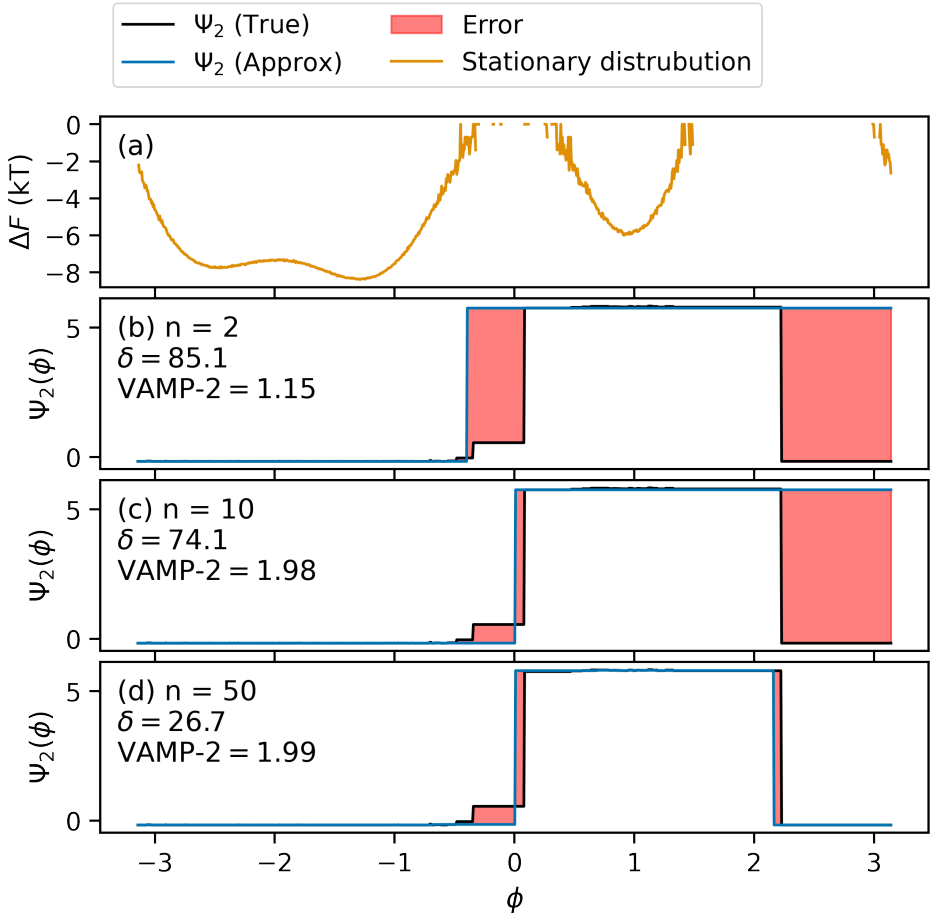


FIGURE 4.10: DISCRETIZATION ERROR OF THE SECOND RIGHT EIGENFUNCTION OF ALANINE DIPEPTIDE AS A FUNCTION OF THE NUMBER OF MICROSTATES. The feature used is the ϕ backbone torsion. Panel (a) shows the free energy along this feature. Panels (b), (c) and (d) show the normalized MSM right eigenfunction (blue line) estimated $n = 2, 10$ and 50 cluster centers respectively. This is compared with the same eigenfunction estimated with $n = 500$ cluster centers (black line). The red shaded area represents the difference between the two eigenfunctions. The discretization error, labelled δ is the integral of the red area. The $VAMP-2$ score is also labelled for comparison.

using discrete microstates and is given by [64]:

$$\delta_i \equiv \left\| \Psi_i(\mathbf{z}) - \hat{\Psi}_i(\mathbf{z}) \right\|_{\pi,2} = \left(\int_\Omega d\mathbf{z} \pi(\mathbf{z}) (\Psi_i(\mathbf{z}) - \hat{\Psi}_i(\mathbf{z}))^2 \right)^{1/2} \tag{4.21}$$

Here $\mathbf{z}$ are the coordinates of state space (in this case the ϕ dihedral angle), π is the stationary distribution and the integral runs over all of the state space, Ω. The integrand is the difference between the true normalized eigenfunction Ψ and approximate eigenfunction $\hat{\Psi}$. As the number of microstates decreases, $\hat{\Psi}_i$ is not fine-grained enough to capture variations in Ψ_i, resulting in large values of δ_i. In the language of statistical learning theory [150], a model with errors due to insufficient flexibility in the model definition, in this case too few microstates, is in the high "bias" regime of the "bias-variance" trade-off.

In contrast, the response is flat for the ϕ and ψ and the RMSD features, which is not expected. This could be because the values of n sampled were not low enough to show discretization error, because of some feature of the simulation data, or due to some error of calculation. In order to understand this, a more detailed investigation of of the relationship between the number of cluster centers, n, the eigenvalue discretization error, δ, and the VAMP-2 response was carried out. This is shown for the ϕ torsion feature in figure 4.10. Panel (a) shows the empirical free energy along the ϕ torsion in order to provide a point of reference for the remaining panels. The truncation of the free energy around the values of $\phi \simeq 0,\ 2$ radians is due to the low temporal resolution of the MD trajectories (each frame is separated by $1\,\mathrm{ps}$). This truncation is shown in the figures of references [98, 327–331] which also use this data.

Panels (b) - (d) show the difference between the 'true' (black solid line) and approximate eigenfunctions (blue solid line) for the slowest relaxation process, Ψ_2. This is the processes which takes the system from the free energy basin on the left-hand side (centered around $\phi \simeq -2$ radians), to the minima on the right hand side ($\phi \simeq +1$ radians). The 'true' eigenfunction was taken as Ψ_2 estimated with $n = 1000$ basis states using the ϕ feature (the shape of the eigenvector changed only slightly over the range $[100, 1000]$). Although the true dominant eigenfunction requires both the ϕ and ψ torsion angles to be described exactly, for the purposes of seeing the effect of n on the discretization error, this definition suffices. As n increases from 2 to 10 to 50, δ decreases from 85.1 to 74.1 to 26.7 (this is the sum of the red shaded area) while the VAMP-2 response increases from 1.25 to 1.98 to 1.99. For this feature, and likely for the other one-dimensional features (ψ, RMSD), the largest decrease in VAMP-2 occurs below $n = 10$ which explains why a drop in VAMP-2 response with decreasing n is not observed.

Second, the response for all features for $n > 100$ is constant. This due to a number of possible factors. First is the large volume of MD simulation data used to train the MSMs. The discretization error will eventually become negligible for all of the slow eigenvectors used in the VAMP-2 score as n increases. As already mentioned, this explains the rapid increase in the response for $n < 100$. As n increases the statistical uncertainty in the elements of the estimated transition matrix will increase and the model enters the high variance regime of the "Bias-variance" trade-off. However, with the

large volume simulation data the number of observations ($750 \times (1000-9) = 743\,250$ pairs of observed transitions) is comparable to the degrees of freedom for a reversible MSM ($^1/_2 n(n-1)+n-1 = 500\,499$ for $n = 1000$), and so the high-variance regime might be at $n > 1000$. In other words, to see the expected decrease in model quality with n, values much larger than 1000 would need to be used. Second, the temporal resolution of the trajectories was not high enough to resolve all the slow relaxation processes stipulated in the VAMP-2 score and so increasing n did not increase the accuracy of the eigenvectors. This is demonstrated in panel (a) of figure 4.10 where the truncation in the free energy surface indicates the upper limit of the resolution.

This flat behaviour is in contrast to other studies [64, 94, 97] where the size of the microstates is important for optimising the MSM. However, with these studies indicator basis function were used whose definition are not dependent on the data. This points to the possibility that for measuring the slow processes, the k-means method which adapts the definition of the state to the data at hand, may not have a practical upper limit on the number of cluster centers (below the pathological limit of the total number of observations).

The third feature of the response surface is the large uncertainty of response surface for $n \leq 20$ which is a result of the comparatively sparse sampling in this region. This was because all sampling was done without prior logarithmic warping, which would have placed more important on small values of n and increased the density of samples in this region.

4.3.1.4 Practical implications

A Gaussian process regression model can be used to model the VAMP-2 response of a Markov state model to both integer valued hyperparameters (number of microstates) and categorical variables (the protein feature). A fully multiplicative kernel of the type in equation 4.13 can by used to model multiple hyperparameters, while different kernel functions can be selected by choosing the model with the best combination of standardized log loss and mean square error. Although this is not useful in isolation, this is an important stepping stone to both measuring hyperparameter relevance and optimising MSMs.

4.3.2 Hyperparameter relevance

Figure 4.11 shows the relevance of the hyperparameters of an MSM of alanine dipeptide. The box plots show the variability in the estimates. This variability is due to the fact that the hyperparameter trial observations are not perfectly predicted by the Gaussian process regression model. The color coding reflects the fact that the levels of the protein feature, χ, which are categorical variables (shown in blue), are to be interpreted differently from the integer valued, number of microstates, n (shown in orange). However, overall, the values for each hyperparameter are less than 1, indicating the shape of the response surface is largely flat. The implications of this are explored below. The median and $95\,\%$ credible intervals of the results in this figure are tabulated in table 4.3.

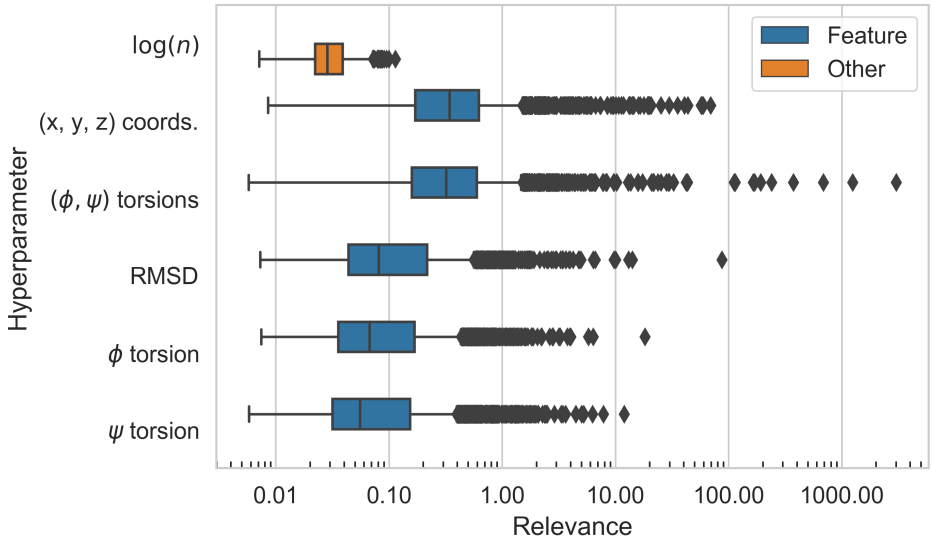


FIGURE 4.11: RELEVANCE OF THE HYPERPARAMETERS OF ALANINE DIPEPTIDE. The distribution of the parameters of the response surface (shown in figure 4.9) were estimated using MCMC. The relevance of the features (levels of χ) are shown in blue, labelled 'Feature'. The relevance of the log-transformed number of cluster centres, n is shown in orange (labelled 'Other').

Hyper-parameter	Median	95% C.I.
$R_{(\phi,\psi)\text{ torsion}}$	0.321	0.020-4.456
$R_{(x,y,z)\text{ coords.}}$	0.344	0.024-5.572
$R_{\phi\text{ torsion}}$	0.068	0.015-1.176
$R_{\psi\text{ torsion}}$	0.056	0.013-1.327
R_{RMSD}	0.081	0.016-1.406
$R_{\log(n)}$	0.029	0.013-0.063
η	2.518	1.141-5.530
σ_n	0.006	0.006-0.007

TABLE 4.3: POSTERIOR DISTRIBUTION OF GP HYPERPARAMETERS. Median and 95% credible intervals for the kernel hyperparameters of the alanine dipeptide response surface estimated using MCMC. The length-scale parameters in 4.13 are re-written here as relevances.

Figure 4.11 clearly shows the low relevance of the number of microstates (shown in orange, median value 0.029, $95\,\%$ C.I. $[0.013–0.064]$, table 4.3). The relevance of continuous hyperparameters determines how sensitive the model response is to changes in that hyperparameter. The more relevant a hyperparameter, the greater the change in model response is to a change its value. Such a low value indicates that the number of microstates only has a negligible effect on the VAMP-2 response of the MSM. This is also evident from looking at the response surface itself, figure 4.9 which shows the response flat with respect to changes in n. The practical implication is that it is not necessary to optimize the number of microstates in an MSM when using k-means clustering (the method used to create the microstates for the models fit here). However, for very small values, the value of n does determine the VAMP-2 score, as was discussed in the previous section. This means that the assumption of stationarity does not hold as correlation in the response for low values of n is different to that for large values of n.

Previous work [306] calculated the relevance of continuous hyperparameters only. This work extends the idea by considering categorical hyperparameters, namely the protein feature χ. The levels of χ are the various protein features of the hyperparameter search space, listed in table 4.2. The ϕ and ψ angle features are shown in figure 4.2. The (x, y, z) coordinates feature is the atomic coordinates of the non-hydrogen atoms, also shown in this figure. The RMSD feature is the root mean square deviation of the non-hydrogen atoms relative to an reference structure.

The relevance of χ determines the amount of information sharing between the parts of the response surface with different values of χ. In order to understand this, it will be useful to contrast it with the relevance of a continuous or integer-valued hyperparameters. The relevance of n determines the covariance of the response to changes in n *within the same feature*. This can be seen from the equation 4.13 and making use of the fact that all stationary kernel functions, $k(x, x') = 1$ for $x - x' = 0$:

$$\begin{aligned} k^{tot}(\mathbf{x}, \mathbf{x}') &= k\big((1,0,0,0,0,n),(1,0,0,0,0,n')\big) \\ &= \eta^2 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot k_M(n, n'; R_n) \\ &= \eta^2 \cdot k(n, n') \end{aligned}$$

Here the kernel functions have been re-written with the relevance, R, instead of the length-scale l as a hyperparameter. The median relevance of $\log(n)$ is equal to 0.029 which implies that for change $n = 10$ and $n = 1000$ (a change of 1 on the normalized scale) the covariance will be $\eta^2 k_{M-52}(0, 1; 0.029) \simeq 0.99\eta^2$. In other words, the response will be independent of the n as already noted. In contrast, the relevance of the dummy coded categorical variable determines the amount of information sharing between different levels of that variable [350]. Between a high relevance protein feature and all other features, there is little information sharing; between low relevance protein features there is a large amount of information sharing. To see this, consider the covariance between

points at n and n' on two different features, χ_1 and χ_2:

$$\begin{aligned} k^{tot}(\mathbf{x},\mathbf{x}') &= k\big((1,0,0,0,0,n),(0,1,0,0,0,n')\big) \\ &= \eta^2 \cdot k_M\big(1,0;R_{\chi_1}\big) \cdot k_M\big(0,1;R_{\chi_2}\big) \cdot 1 \cdot 1 \cdot 1 \cdot k_M(n,n';R_n) \\ &= \eta^2 \cdot k_1 \cdot k_2 \cdot k(n,n') \end{aligned}$$

If either R_{χ_1} or R_{χ_2} is large then $k_1 \cdot k_2 \simeq 0$ and there will be no correlation between n on feature χ_1 and n' on feature χ_2. If both R_{χ_1} and R_{χ_2} are small then $k_1 \cdot k_2 \simeq 1$ and covariance between n on feature χ_1 and n' on χ_2 will be similar to the covariance between n and n' on the same feature.

Figure 4.11 clearly shows that features for alanine dipeptide are all low relevance. This means the model response to n will be similar for the different values of χ. Even between the two highest relevance features (ϕ,ψ) (median relevance $= 0.321$, 95% C.I. $[0.020\text{–}4.456]$) and (x,y,z) (median relevance $= 0.344$, 95% C.I. $[0.024\text{–}5.572]$) the covariance between n and n' on these two features is only altered by $k_1 \cdot k_2 = 0.91 \cdot 0.92 \simeq 0.83$.

4.3.2.1 Practical implications

Using an estimated response surface, the relevance of hyperparameters of an MSM can be estimated. Time should be spent optimising high relevance continuous or integer-valued hyperparameters as small changes in their value imply large changes in model response. Similarly, high relevance hyperparameters should be chosen for sensitivity tests, which test the robustness of scientific conclusions to modelling choices (this will be explored in chapter 6). For categorical hyperparameters such as the protein feature, χ, the story is subtly different. A group of low relevance values of a categorical hyperparameter can all be treated similarly with respect to the other hyperparameters. With respect to work here - the optimum value of n is then similar for all values of protein feature. The relevance also provides a useful guide for visualising multidimensional response surfaces, although this was not investigated here: plotting high relevance hyperparameters can take priority over low relevance hyperparameters, without sacrificing important features in the response surface. This idea will be taken up in chapter 6.

4.3.3 Optimization

The maximum of the response surface of alanine dipeptide, estimated using the $N = 500$ randomly sampled hyperparameters with a GP regression model, is $\mu = 3.318 \pm 0.004$ which corresponds to using the (x,y,z) coordinates feature with $n = 762$ cluster centers. The maximum of the response surface *at the trial values* gives the optimum hyperparameters incorporating uncertainty and making full use of all the trial information. Given the simplicity of the response surface, visual inspection of figure 4.9 was deemed sufficient to confirm that no more sampling was necessary to locate the maximum.

However given this large amount of sampling may not be practical in settings where estimating an MSM is more intensive due to larger amounts of data, a useful question to ask is then:

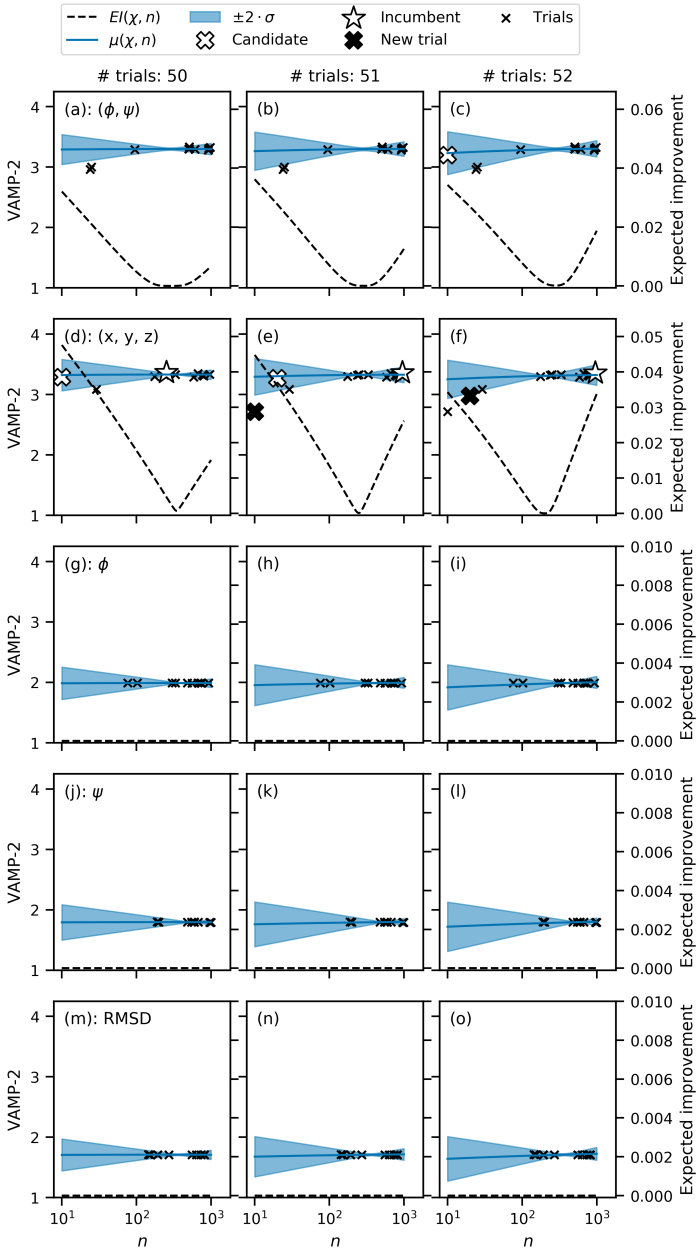


FIGURE 4.12: BAYESIAN OPTIMISATION. Each column shows the response surface at three sequential points in the Bayesian optimisation procedure and each row corresponds to a different feature. The vertical axis is the MSM response, and the horizontal axis the number of microstates on a logarithmic scale. The blue line and shaded area show the response surface (mean and uncertainty respectively) estimated using the hyperparameter trial data set, shown as small black crosses. The white star shows the incumbent and the black dashed line shows the expected improvement (right hand scale - note different vertical scales for each row). The white cross shows the candidate hyperparameter, i.e., the maximum of the acquisition function. The large black cross show the actual value of the hyperparameter trial that was suggested in the previous column.

		N_{total}		μ		σ		χ		n	
N_{seed}	#	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
0	1	500		3.318		0.002		(x, y, z)		762	
30	1	30	40	3.302	3.338	0.004	0.192	(ϕ, ψ)	(x, y, z)	577	969
	2	30	40	3.318	3.233	0.005	0.086	(ϕ, ψ)	(x, y, z)	540	133
	3	30	40	3.076	2.947	0.557	0.420	(ϕ, ψ)	(ϕ, ψ)	88	10
	4	30	40	3.065	3.315	0.553	0.132	(x, y, z)	(x, y, z)	627	1000
	5	30	40	3.313	3.258	0.005	0.194	(x, y, z)	(ϕ, ψ)	968	684
50	1	50	60	3.330	3.337	0.012	0.032	(x, y, z)	(x, y, z)	251	333
	2	50	60	3.306	3.338	0.022	0.040	(ϕ, ψ)	(ϕ, ψ)	540	540
	3	50	60	3.309	3.327	0.013	0.013	(x, y, z)	(x, y, z)	176	670
	4	50	60	3.307	3.318	0.005	0.004	(ϕ, ψ)	(x, y, z)	634	1000
	5	50	60	3.308	3.327	0.004	0.058	(x, y, z)	(x, y, z)	390	314

TABLE 4.4: MSM HYPERPARAMETERS FOR ALANINE DIPEPTIDE PRE- AND POST-BAYESIAN OPTIMISATION. Each row represents a BO experiment, seeded with N_{seed} randomly sampled hyperparameters trials. Five iterations of optimisation were run with $N_{\text{seed}} = 30,\ 50$ (labelled #1,2 etc.). The number of optimisation steps is equal to the difference in the pre/post value of N_{total}. The optimum of the response surface estimated with all the trial data ($N_{\text{total}} = 500$) is included even though it was not optimised using BO. Each column is a variable or outcome with values associated with the optimum value of μ, before and after BO.

"Can the optimum of the response surface be discovered with less hyperparameter trials using Bayesian optimisation?". Before answering this question it is worth explaining how the Bayesian optimisation process works in practice. The Bayesian optimisation process for three consecutive steps is demonstrated in figure 4.12. The first column (panels (a), (d), etc.) shows the response surface estimated using the hyperparameter trial data set with $N = 50$ trials. The blue line and shaded area in panel (d) are the response surface with $\chi = (x, y, z)$. The white star is the incumbent - the maximum of the response surface, across all features but only evaluated for $\mathbf{x} \in \mathscr{D}_{50}$. The expected improvement is shown as a dashed line and its maximum corresponds to the point on the response surface denoted by the white cross. This value, $\mathbf{x}_{50} = \big(\chi = (x, y, z), n = 10\big)$ is the candidate and is evaluated in the next step. In the second column (panels (b), (e) etc.) $\mathbf{x}_{50}$ has been evaluated and the result is shown as a filled black cross. Its value is much smaller than expected, although the new response surface adapts poorly to this information - the mean response does not pass through this new point. This process is repeated in the next column (panels (c), (f) etc.). The white cross in panel (e) is evaluated and shown as a black cross in panel (f) (the candidate in panel (e) is not at the maximum of the acquisition function because the maximum was the same as the previous step).

Bayesian optimisation was performed starting with a small set of randomly sampled hyperparameter trials (pairs of hyperparameters and VAMP-2 responses) to initiate the process. This simulates the more realistic situation of large amounts of simulation data prohibiting exhaustive searches of

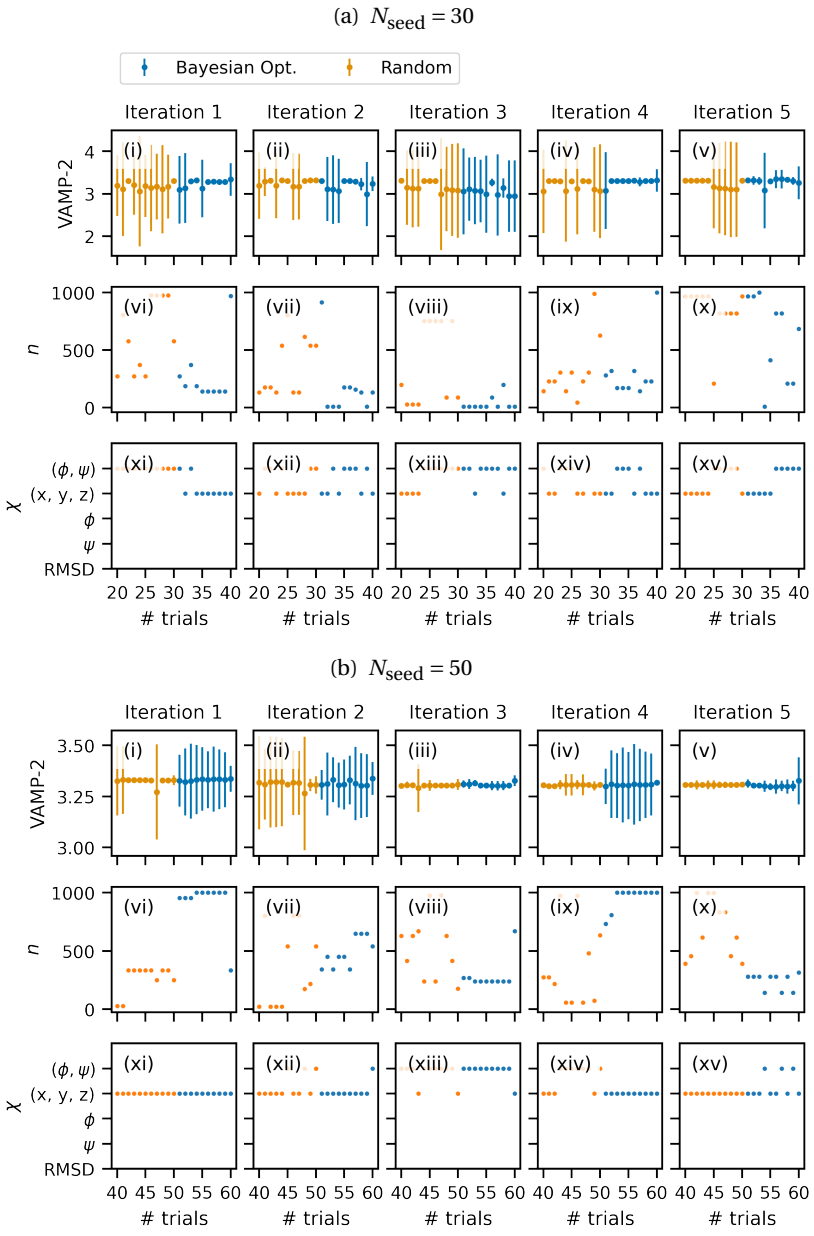


FIGURE 4.13: BAYESIAN OPTIMISATION TRAJECTORIES OF ALANINE DIPEPTIDE. Shown are five different random subsets ('iterations') of the total hyperparameter trial data, each separately optimised, seeded with 30 hyperparameter trials or 15 observations per predictor (panel (a)), and 50 hyperparameter trials or 25 observations per predictor (panel (b)). The orange values are the trajectories calculated from random sampling, the blue values are the Bayesian optimisation trajectories. The first row (sub-panels (i) - (v)) are the VAMP-2 response, the second row (sub-panels (vi) - (x)) show the accompanying number of cluster centres, and the third row (sub-panels (xi) - (v)) are the accompanying feature.

the hyperparameter space. Specifically ten steps of Bayesian optimisation was performed on five, randomly sampled, subsets of the full hyperparameter trial data set with sizes $N_{\text{seed}} = 30$ & 50. The input warping and kernel function used in the response surfaces for all of the Bayesian optimisation experiments were the same as those used on the full trial data set, discussed in the previous two sections. In principle, these modelling choices should be determined independently for each data set but, given the simplicity of the response surface, it was deemed unnecessary.

The results of the optimisation are shown Figure 4.13 by plotting the *optimisation trajectories*. Figure 4.13(a) shows the optimisation trajectories after seeding with $N_{\text{seed}} = 30$ trials and figure 4.13(b) after seeding with $N_{\text{seed}} = 30$ trials. The 10 steps of Bayesian optimisation are shown in blue (horizontal axis values: $N_{\text{seed}} \rightarrow N_{\text{seed}} + 10$) and for comparison the figure also shows, in orange, the trajectory calculated using randomly sampled trials (horizontal axis values: $N_{\text{seed}} - 10 \rightarrow N_{\text{seed}}$). Sub-panels (i) - (v) show how the value of the incumbent varies with the number trials in the trail data set, the orange values contain only randomly selected hyperparameters, the blue values contains a mix of BO selected and randomly selected trials. Panels (vi) - (x) show the values of n and panels (xi) - (xv) show the values of the χ associated with the incumbent.

With $N_{\text{seed}} = 30$ the incumbent trajectories of iterations 2, 3, & 5 decreased after starting BO. For all iterations seeded with $N_{\text{seed}} = 50$ trials the incumbent trajectories remained constant or increased. In both cases the optimal values of χ oscillated between (ϕ, ψ) and (x, y, z), while the optimal value of n did not converge to a single value across separate iterations. A value of $N_{seed} = 50$ or 25 observations per predictor were therefore deemed tentatively appropriate.

There are a number of observations of the optimisation trajectories which reflect on the alanine dipeptide response surface described in section 4.3.1 and the usefulness of Bayesian optimisation for this system. First, the incumbent trajectories clearly show that Bayesian optimisation does not increase the value of the incumbent by a significant amount. This is a reflection of the simple nature of the response surface and the irrelevance of the number of cluster centres. Second, the response surface (in the search space domain tested) is bimodal with peaks at $\chi = (\phi, \psi)$ torsions and (x, y, z) coordinates. This is reflected in the clear lack of substantive difference between the final values of μ listed in table 4.4 for these two values of χ. Third, the almost complete irrelevance of n as a hyperparameter is clearly shown in figure 4.13(b) sub-panels (vi) to (x) in which the final values of n, $n \simeq 1000$, 500 & 100 show no clear difference in the value of the incumbent. Fourth, it is possible that with more optimisation steps it could be possible to arrive at the maximum of the response surface with fewer seed trial observations. While this is a possibility, the fact that Bayesian optimisation is an inherently serial algorithm, while random sampling is embarrassingly parallel, it was considered more wall-time efficient (if not CPU-time efficient) to err on the side of more random seed trial data and fewer optimisation steps.

4.3.3.1 Practical implications

Bayesian optimisation may not be necessary for optimising MSMs. The nature of the response surface of alanine dipeptide is such that no improvements could be made over after randomly sampling, and in some cases made the measured response worse. However, Bayesian optimisation does provide a useful check of convergence of the response surface: the Bayesian optimisation algorithm explored the hyperparameter space in the regions most likely to show improvement ($\chi = (\phi, \psi)$ torsions and (x, y, z) coordinates) but could not improve on the VAMP-2 score.

4.4 Conclusions

This chapter introduced the use of response surfaces and Bayesian optimisation for understanding and optimizing the hyperparameters of MSMs. A GP model proved a satisfactory statistical model for estimating the response surface of alanine dipeptide with two predictors χ and n. Using the MSLL and SMSE metrics, a Matérn 5-2 kernel and logarithmic warping of the hyperparameter n produced the best fit to the data. The logarithmic warping was necessary to make the stationary assumption of the GP more plausible. While GPs are usually used with continuous predictors, the use of dummy coding was demonstrated to be effective in incorporating the peptide feature, χ, as a categorical predictor. The hyperparameter relevance, the inverse of the characteristic length-scale of the kernel, was shown to reflect the importance of each MSM hyperparameter in determining the MSM response. For the non-categorical hyperparameter, n, the low relevance was a reflection of the near flat response of the MSM to changes in n. For the categorical predictor, the protein feature χ, the low relevances of each feature was a reflection of how similar the response surfaces were *conditional* on the value of χ. Two BO experiments, each consisting of five iterations of the Bayesian optimisation procedure were performed. It was found that seeding the algorithm procedure with 30 hyperparameter trials led to the Bayesian optimisation procedure deteriorating rather than optimising the hyperparameter surface. Seeding with 50 trial observations prevented the deterioration in the incumbent, however, it did not improve the incumbent either. The hyperparameters selected by Bayesian optimisation over each iteration were not consistent.

The main limitation of this work is that the response surface of alanine dipeptide was too simple to provide an adequate test of the methods investigated here. The response surface could be easily visualised in its entirety and could be approximated as a constant value for each feature, with the two best performing features having equal response values. Thus the hyperparameter relevance calculations did not provide any more insight beyond what could be gathered from visual inspection of figure 4.9.

Bayesian optimisation did not provide any useful reduction in computational effort when optimising this response surface because of the shape of the response surface. Thus random sampling was easily able to pick out the optimum hyperparameters. The possible reasons for the flat response with respect to n is the large volume and low temporal resolution of the trajectories. In

addition, the type of clustering algorithm is known to have a large effect on the quality of an MSM [246] and the effect of this has not been investigated in this work. As a result, Bayesian optimisation was only shown to be *as good* as randomly selecting hyperparameters and fitting a response surface but not better. However, it did provide a convergence check on the randomly selected variables, so in that limited sense it did provide an reduction in computational effort. In addition, the model selection criteria for selecting the kernel and input warpings was simple and easily implemented but ultimately ad-hoc. A plan to address these limitations in future work is laid out in the conclusions, chapter 7.

However, as the majority of these limitations were dependent on the system studied, they do not rule out their practical benefits for other systems. First, a Gaussian process was able to be fit to estimate the response surface, despite the fact that the data did not meet the stationarity assumption, required by the kernel. Second, the hyperparameter relevance was consistent with the data. This work demonstrated an interpretive extension of hyperparameter relevance for categorical features. This has benefits when deciding on how to further optimise or visualise the response surface, or when designing sensitivity tests. Third, while Bayesian optimisation was not able to optimize the response surface, it did provide a useful check of convergence of the optimum of the response surface. A further test of these methods will be presented in chapter 6, where in addition the use of hyperparameter relevance for visualising multidimensional response surfaces and designing sensitivity tests will be explored further.

CHAPTER 5

METASTABLE STATE SELECTION FOR HIDDEN MARKOV MODELS

Symbol	Definition
g	Number of hidden states in a HMM.
n	Number of observed states in a HMM.
$\tilde{\mathbf{T}}$	$g \times g$ HMM transition matrix.
$\mathbf{E}$	$g \times n$ emission matrix. E_{ij} is the probability of observing a state j given an hidden state i.
$\tilde{\boldsymbol{\pi}}$	Stationary distribution of hidden states
$\{s_t\}$	Trajectory of observed states
$\{h_t\}$	Trajectory of hidden states
$\mathbb{P}(\{s_t\}\|M)$	Integrated or marginal observed likelihood: the probability of observing $\{s_t\}$ given the model M.
$\mathbb{P}(\{(s_t,h_t)\}\|M)$	Integrated or marginal complete-data likelihood: the probability of observing both $\{s_t\}$ and $\{h_t\}$ given the model M.
$\mathscr{L}\left(\tilde{\mathbf{T}},\mathbf{E}\|\{s_t\}\right)$	HMM likelihood.
θ	The HMM parameters, $\tilde{\mathbf{T}}$ and $\mathbf{E}$
$\hat{\theta}$	Maximum likelihood estimates of θ
$\mathbf{M}$	Membership matrix. M_{ij} is the probability that a given observed state i is a member of hidden state j.
$H(s_t;\mathbf{M})$	The information entropy associated with observed state s_t.
$EN(\mathbf{M})$	Classification entropy - the sum of the $H(s_t;\mathbf{M})$ over a whole trajectory data set.

TABLE 5.1: IMPORTANT SYMBOLS USED THROUGHOUT THIS CHAPTER.

5.1 Introduction

Chapter 4 demonstrated using response surface methods and Bayesian optimisation to arrive at an optimal MSM. Typically MSMs are created with 100s of basis states, or microstates, which are geometrically similar so that they describe with a high degree of precision the eigenvectors of $\mathbf{T}$. These eigenvectors describe dynamic processes such as loop opening and closing (e.g., in triosephosphate isomerase [6] described in chapter 1) or rotations about dihedral angles (e.g., in alanine dipeptide, chapter 4). While the microstate basis can provide quantitative agreement with experimental lifetimes and rates [351], descriptions in terms of 100s of states, which only differ by small changes in atomic configuration, are cumbersome and difficult to interpret. It is common practice to instead group "similar" (this will be made precise below) microstates into a small number of meaningful and interpretable *macrostates*. To continue the loop opening example of triosephosphate isomerase (TIM), the authors of reference [6] created an MSM with 100 microstates, each one of which differed slightly along 120 different interatomic distances. Rather than stop at this description, they showed that these microstates could be grouped into just two or three macrostates which provided a more meaningful description in terms of the loop being open or closed.

When grouping or *coarse-graining* microstates into macrostates there are three questions that need to be answered:

1. How are the microstates mapped to the macrostates? i.e., for each microstate which macrostate does it get assigned to?
2. Given a mapping, what are lifetimes of, and transition rates between, the macrostates? In other words, what is the transition matrix $\widetilde{\mathbf{T}}$ in this new basis?
3. Model selection: How can competing models be compared and selected? This includes selecting among different coarse-graining methods and selecting the number of macrostates.

It may also be useful to ask what is the purpose of the performing the coarse-graining? Is it to classify states crisply into disjoint macrostates? Or is it to produce the most accurate description the kinetics in a coarse-grained basis? These two purposes require different model selection criteria and may affect the optimum coarse-graining scheme.

5.1.1 Assigning microstates to macrostates

There are two main approaches to coarse-graining and clustering in general, which have already been briefly discussed in section 1.7 of the introduction to this thesis. First, geometric clustering can be used to assign snapshots from molecular dynamics (MD) trajectories to microstates based on their geometric similarity, for example the root mean square deviation from a reference structure. The second approach is kinetic clustering where clustering is performed by explicitly grouping frames into states based on their kinetic properties. Kinetic properties of a particular atomic configuration are determined by its location on the free energy landscape. In order to transition from one region

to another a barrier, which combines enthalpic (potential energy) and entropic terms, must be surmounted [102]. Thus kinetic clustering amounts to grouping molecular configuration which lie in the same minima on the free energy surface into the same state (defining clusters which lie near or on transition states is also important and will be described below). Discrete path sampling is a method of kinetically clustering atomic configurations [101]. To do this the free energy surface is estimated by first finding minima and saddle points (transition states) in the *potential energy* surface and then adding an estimate of the entropic contribution. However, this does not scale to large systems because finding potential energy minima results in large energy fluctuations with systems with large numbers of degrees of freedom (for example with explicitly solvated systems) [63]. For larger systems, the MSM approach is to approximate kinetic similarity with geometric similarity. The preprocessing steps of projecting onto relevant molecular features and time-lagged independent component analysis (described in chapter 2), is to create a set of variables which allow identification of geometric distance with the kinetic distance [90]. So if the distance between frames as measured in these new variables is small, then they can be assumed to have similar kinetic properties.

Coarse-graining microstates into macrostates is conceptually similar to clustering MD frames or atomic configurations into microstates, except that the time or distance scale has been enlarged because of the questions being asked. Understanding loop motions, which occur nanosecond and microsecond timescale, does not require resolution of the rotameric states of individual methyl groups, which transition on the picosecond timescale [352].

There have been a wide variety of methods developed for coarse-graining microstates into macrostates. The majority of these methods create macrostates corresponding to the same free energy minima. This means that two microstates, a_i and a_j, in the same macrostate, A, interconvert more rapidly than to microstates b_k in a different macrostate, B. This is possible when there are differences in the timescales of the processes that cause transitions between A and B versus the transitions within A (or $B, C, \ldots$). Under these circumstances the macrostates are said to exhibit metastable dynamics [169]. Coarse-graining can also be used to find macrostates which lie on transition regions between metastable states, known as transition macrostates [157].

The most straight-forward method of coarse-graining into metastable macrostates is when the slow processes can be easily identified with a protein feature. In reference [71] a 200 state MSM was constructed to model allosteric modulation in the enzyme cyclophilin A. The slow process was identified with a change in the χ_1 angle of the Phe113 residue. Coarse-graining could then be done 'by hand' by assigning microstates into either of two macrostates based on whether $\chi_1 = \pm 60°$. When the number of relevant degrees of freedom makes this infeasible, algorithmic methods are necessary. The first method for explicitly coarse-graining the microstates of an MSM was Perron cluster cluster analysis (PCCA) [158] and its subsequent 'robust' alternative PCCA+ [159] (which improved the method under low-data conditions). Each eigenvector of the transition matrix $\mathbf{T}$ represents a dynamic process (e.g., loop opening) and the associated eigenvalue is related to its timescale. The relative sign of microstates in this eigenvector denotes how the process evolves over

time: microstates with a negative sign transition to states with positive sign and vice versa. PCCA uses this sign structure to coarse-grain states: microstates with the same sign get grouped into the same macrostate (see the description of PCCA+ in section 2.4.2). Some recent examples include using PCCA+ to coarse-grain: the 100 microstates of loop 6 into 'open' and 'shut' macrostates in TIM [6]; 237 microstates of the protease trypsin into seven macrostates each of which exhibited different binding affinities to the inhibitor benzamidine [160]; 500 microstates of DNA quadruplexes into five metastable states in order to test the accuracy of different molecular mechanics force-fields compared to X-ray crystallography experiments [161].

PCCA+ is also used in the hierarchical Nyström exstension graph method (HNEG) [162]. In the HNEG method, instead of the whole microstate basis, the transition matrix $\mathbf{T}$ is approximated with a subset of only the most well sampled states (the Nyström method is a method for approximating matrices with an appropriate submatrix). PCCA+ is used to coarse-grain the submatrix and the remaining microstates are grouped based on their transition probabilities to each macrostate. HNEG has been used to understand the binding interface of the calcium sensor calmodulin [240]. The Bayesian agglomerative cluster engine (BACE) [163] uses Bayesian hypothesis testing to test whether a given microstate belongs in a given macrostate. BACE has been applied to modelling the loop dynamics of MDM2 (mouse double minute 2 homologue - a cancer drug target), coarse-graining a 2000 state MSM into 150 macrostates. HNEG and BACE are similar because they first assign well-sampled microstates, resulting in coarse-grainings which are more robust when including poorly sampled microstates. The most probable path (MPP) [164] assigns microstates to the same macrostate if they occur on the sequence of microstates. This sequence is constructed by starting with a microstate i, the next state has the highest transition probability from i (i.e., $\mathrm{argmax}_j\, T_{i,j}$) and so on. MPP has been used to assign a 12000 microstate description of the villin headpiece protein into five macrostates representing one unfolded, two intermediate and two folded states [353]. The renormalisation group is a technique for making links between phenomena at different scales e.g., the link between microscopic atomic interactions and macroscopic phase transitions [354]. This technique has been applied to MSMs [165] to derive a coarse-graining method which was used by the authors to reduce a 100 microstate description of the bovine pancreatic trypsin inhibitor to three metastable folded states. A conceptually different approach was taken with minimum variance cluster analysis (MVCA) [166]. Row i of the transition matrix $T_{i,j}$ defines a probability distribution over all other microstates j. Two microstates i and k are similar if they have similar distributions $T_{i,j}$ and $T_{i,k}$. MVCA hierarchically clusters microstates based on a measure of the similarity of the two distributions. The authors used this to compare the number of metastable states of the fast folding protein chignolin with two different molecular mechanics force-fields. They also extended this approach to cluster not just rows of a single transition matrix but also whole transition matrices. They used this technique to highlight the differences and similarities between MSMs of simulations of chignolin using different combinations of force-fields and water models.

Projected Markov models include observer operator models (OOMs) [167] and hidden Markov

models (HMMs) [169]. These are dynamical models which directly incorporate a microstate-macrostate structure into their definition. PMMs assume Markovian dynamics in an unobserved state space and then specify a projection onto observable variables. OOMs and HMMs are different approximations to this process. HMMs have been described in chapter 2 but to summarise: HMMs specify a Markov process in hidden states (macrostates), while in each hidden state the system 'emits' an observed microstate. Each macrostate emits with a different set of probabilities to each microstate. This emission distribution defines the coarse-graining. HMMs are accurate representations of PMMs assuming the unobserved dynamics are metastable and the emission distributions do not overlap. OOMs on the other hand do not require non-overlapping emission distributions and are therefore considered the more general model [167]. HMMs have proved the more popular model and have been used to model protein-protein association dynamics [174], ligand binding [173, 182–184, 186], enzyme-substrate dynamics [175, 177, 180], nucleic acid folding dynamics [178, 179], ion channels [181], allosteric effects [185] and to aid in enzyme design [176]. This is likely due to their attractive properties for understanding conformational dynamics. First, there is a clear one-to-one relationship between the elements of the model and the intuition about the conformational dynamics of biomolecules: the rapidly inter-converting configurations correspond to the observed states of the HMM, while the metastable states correspond to the hidden states of the HMM [169]. Second, unlike MSMs, the dynamics of the observed states (or microstates in the language of MSMs) are not required to be Markovian in order to recover accurate relaxation timescales. Third, they have been shown to be robust to poor microstate definitions [169].

The previous techniques all focused on creating definitions of macrostates which were metastable, i.e., correspond to minima in the free energy surface. However, transition states are crucial to describing the rates of conformational transitions and increasing the number of microstates around the barrier separating free energy minima increases the accuracy of Markov models [64]. In reference [157] the authors coarse-grain microstates to find both metastable *and* transition-state macrostates. The central idea is to select macrostates which maximize the slowest implied timescale in the macrostate model. This requires a variational bound on the slowest timescale in the coarse-grained picture, which was later proved in reference [355]. The authors went on to apply this to a wide variety of systems: quantum-mechanics/molecular mechanics multiscale simulations of hydrogen abstraction in lipoxygenase, the model helix forming alanine pentapeptide, and two different helix dimers.

5.1.2 Coarse-grained description of kinetics

In order to make the link between the coarse-grained description and observable dynamical properties of the system and new coarse-grained transition matrix is required. In general, defining a new macrostate transition matrix by simply merging microstates will result in a non-Markovian description which fails to correctly propagate the dynamics (i.e., the elements of $\mathbf{p}$ in the master equation equations 1.1 and 1.2) [169, 236]. The authors of reference [236] attempted to define a coarse-

grained transition matrix by minimizing the error between the predicted populations of states in macrostate and microstate basis. As their numerical example illustrated however, their method could result in unrealistic negative transition probabilities. A more successful approach was proposed by Hummer and Szabo [155]. Their method enforced consistency between the microstate and macrostate descriptions of the relaxation times between macrostates. This result was used in the method of reference [157] to identify transition-state macrostates described previously. The Hummer-Szabo method was used in reference [356] to calculate the macrostate transition matrix describing the helix-coil transition in an α-helical peptide. The macrostate transition matrix was derived from a 199 microstate transition matrix after coarse-graining with PCCA+. Later work also showed [156] their method to be equivalent preserving mean first passage times (the average time taken to transition between states) between the microstate and macrostate basis.

For some coarse-graining techniques the construction of a consistent macrostate transition matrix is more straightforward. The renormalization group technique [165] defines simultaneously the micro-macrostate mapping and the macrostate rate matrix and so no further work is necessary. For projected Markov models (HMMs and OOMs) the transition matrix is estimated directly from the data. For HMMs this is done by maximizing the likelihood of the model parameters (the transition matrix elements and the emission distribution parameters) or through Bayesian estimation. In both types of estimation the likelihood function makes the link between the observed trajectories of microstates, the coarse-grained transition matrix, and the assignment of microstates to macrostates. Each observed microstate in a trajectory can be thought of as arising from two processes: emission and transition. At time t the system is in a given macrostate h_t; this macrostate 'emits' a microstate s_t with a given probability E_{h_t,s_t}. The macrostate itself arose because of a transition from an earlier macrostate h_{t-1} with a probability of $\widetilde{T}_{h_{t-1},h_t}$. The likelihood function relates the observed microstates to elements of the matrices $\mathbf{E}$ and $\widetilde{\mathbf{T}}$. The optimal macrostate basis is defined by finding $\mathbf{E}$ and $\widetilde{\mathbf{T}}$ which maximize the likelihood. This is covered in detail in chapter 2 and later in section 5.2.

5.1.3 Model selection

The number of macrostates, g, must be stipulated when coarse-graining a MSM but choosing an appropriate value of g is complicated by two factors. First, the nature of complex systems is that there is a hierarchy of timescales [107, 352] which define multiple different metastable and transition-state coarse-graining. The number of macrostates must therefore be set in relation to a timescale. Even after setting this timescale, none of the methods so far give a definitive answer as to how many macrostates should be specified, however, some give more information to guide the user than others. For example, the Bayesian agglomerative clustering engine [163] hierarchically assigns microstates, initially to a large number of macrostates, which are in turn assigned into larger macrostates until only two macrostates remain. Thus the user can see how states agglomerate and choose the number of macrostates according to their specific analysis objective. The 'distance'

between the macrostates (i.e., the number of iterations before two distinct states are merged) is measured by the Bayesian weight of evidence for merging two states (this is the Bayes factor, discussed below). By requiring stronger evidence for merging two states a more fine grained picture with more macrostates is produced. In the transition-state coarse-graining method of reference [157] the algorithm can by stopped after identification of a single transition-state macrostate or can be repeated, finding more transition-state macrostates, which again produces a more fine-grained picture. A more general method for determining the number of metastable macrostates is to inspect the eigenvalue spectrum of the MSM in the microstate basis. Gaps in the eigenvalues or the implied timescales [169] denote regions where there is a separation in timescales. A large gap between the g and the $g+1$'th timescale supports coarse-graining into g metastable macrostates.

The second complicating factor is the finite amount of data which often leads to gaps in the implied timescale which are statistically indistinguishable from one another [168], making it difficult to determine g. In practice this means that noise in the data could be interpreted as a separate metastable state, a phenomena known as over-fitting [150, chapter 6]. This observation drove the the development of PCCA+ over the original PCCA method [159] and many of the other methods described previously [168].

Deciding on the number of macrostates a given set of simulation data supports is an example of the problem of model selection [150, chapter 6][357]. Chapter 4 was an example of model selection where the choices being decided upon related to the creation of optimum microstates. This focus of this chapter is on the number of metastable macrostates in a coarse-grained Markov model. One approach to determining the number of macrostates is through Bayes factors [189]. The Bayes factor, BF, of two models, M_1 and M_2 relates the posterior odds of two models, given the data, D, to the prior odds of the models [189]:

$$\text{Posterior odds} = \text{Bayes Factor} \times \text{Prior odds} \tag{5.1}$$

$$\frac{\mathbb{P}(M_1|D)}{\mathbb{P}(M_2|D)} = \frac{\mathbb{P}(D|M_1)}{\mathbb{P}(D|M_2)} \times \frac{\mathbb{P}(M_1)}{\mathbb{P}(M_2)} \tag{5.2}$$

$$= \frac{\int \mathbb{P}(D|\theta_1)\,\mathbb{P}(\theta_1|M_1)\mathrm{d}\theta}{\int \mathbb{P}(D|\theta_2)\,\mathbb{P}(\theta_2|M_2)\mathrm{d}\theta} \times \frac{\mathbb{P}(M_1)}{\mathbb{P}(M_2)} \tag{5.3}$$

The integral in the definition of the BF runs over all the potential values of the model parameters, θ, weighted by their prior probability $P(\theta)$. If the prior odds are one, i.e., there is no prior reason to favour one model over another, then the BF is equal to the posterior odds of the two models. If $\mathrm{BF} > 1$ then model 1 is favoured and vice versa. The Bayes factor measures the relative evidence of two models provided by the data [189]. In the case of coarse-graining Markov models for conformational dynamics, the data are the discrete microstate trajectories $D = \{s_1, s_2, ...\} = \{s_t\}$, and the model is, for example a HMM, represented by its parameters $\theta = (\tilde{\mathbf{T}}, \mathbf{E})$ [188]. Practical use of the Bayes factor amounts to calculating the integrated likelihood for each model M_i and selecting the model with the largest value. Using this method, quantitative comparisons of several of the lumping schemes previously cited (excluding HMMs) were compared for a number of benchmark systems

[168]. Bayes factors are attractive as they naturally penalise overly complex models, i.e. models with high-dimensional parameter vectors θ. The reason is that while the likelihood $\mathbb{P}(D|\theta)$ may increase with the dimension θ due to the increased flexibility of the model, the prior probability of any particular θ decreases as it occupies a smaller fraction of this higher dimensional parameter space [189, 358]. However, the main drawback of Bayes factors is the computational effort required to estimate integrals of the type in equation 5.3, which limits the number of models that can be compared.

5.1.4 Hidden Markov models

The proceeding discussion has shown that there exist many different methods for coarse-graining microstates to macrostates (PCCA+, BACE, HNEG etc.). In order to describe the kinetics of the system in the macrostate basis a method for calculating the macrostate transition matrix is required. The method of Hummer and Szabo [155] is one such method which can be applied to different coarse-graining scheme and maintains important properties from the microstate description. Recent work showed that the optimum coarse-grained description can then be derived by maximizing the slowest timescale in macrostate transition matrix [156, 355]. For other methods such the the renormalization group [165] and hidden Markov models [169] the macrostate transition matrix arises directly from application of the method. Choosing the number of macrostates is a problem common to all coarse-graining methods and must take into account the timescale of the problem being investigated and the finite amount of data used to estimate models. Bayes factors are a solution to this problem but require a separate, computationally intensive calculation.

The remaining discussion and the focus of this chapter will be on the use of hidden Markov models for coarse-graining Markov state models. In particular the problem of hidden Markov model selection will be discussed and investigated. The choice of HMMs is motivated because of their established popularity for coarse-graining MSMs. In addition, because of their long-standing use in other fields, there are a wide range of techniques available for model selection which are simpler than the Bayes factor approach. These other techniques and the link to the Bayes factor will constitute the remainder of this introductory discussion.

5.1.5 Classification versus density estimation

Hidden Markov models are a type of finite mixture model [190]. Finite mixture models are used for two purposes (a) modelling the density of observations and (b) classifying observations into meaningful groups [359] both of which can be related to the purposes of coarse-graining MSMs. A coarse-graining is designed to classify observations or microstates into meaningful groups, namely metastable macrostates. Modelling the 'density of observations' in this case means that the macrostate picture should be consistent with equilibrium distribution of microstates and their propagation through time. There is a natural tension between these two purposes. On the one hand more macrostates can model

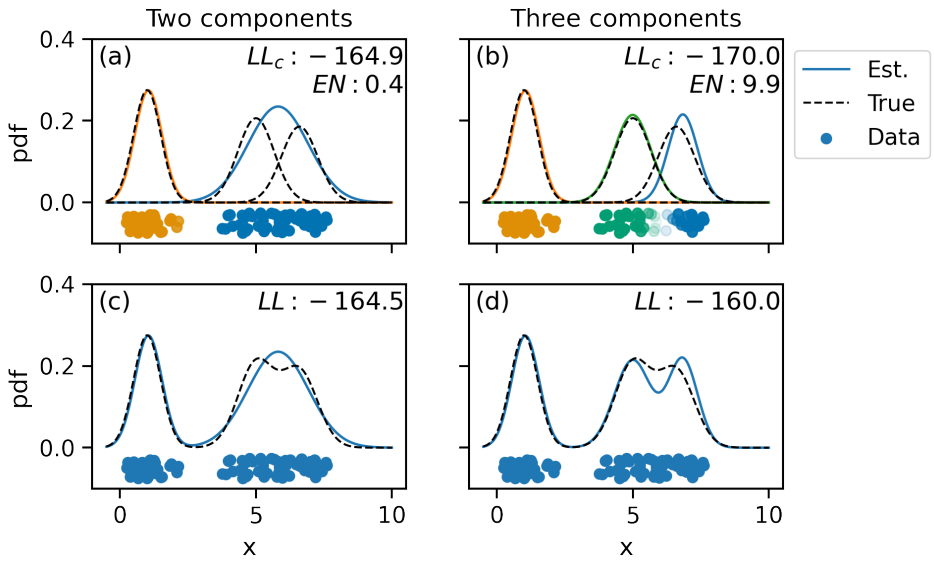


FIGURE 5.1: CLASSIFICATION AND DENSITY ESTIMATION WITH MIXTURE MODELS. Panels (a) and (b) show the classification picture for a two and three component Gaussian mixture model estimated on the same data. The generating (true) densities are shown as black dashed lines, and the generated data shown underneath as coloured discs. The estimated densities are shown as coloured lines. The data have been coloured according to their maximum a posteriori assignment to each estimated component, the transparency of the colour is proportional to the classification entropy - the more uncertain the assignment the more transparent. The label shows the log classification likelihood (LL_c) and entropy (EN). Panels (c) and (d) show the same two models but the densities have been added so as to reproduce the total density of the data. The label shows the log likelihood (LL). The data generating distributions are $\mathcal{N}(1,(1/2)^2)$, $\mathcal{N}(5,(2/3)^2)$, $\mathcal{N}(6.6,(2/3)^2)$, mixed in proportions $\boldsymbol{\pi} = (0.34, 0.34, 0.32)$.

the dynamic processes more accurately but at the expense of losing meaning and interpretability of the macrostates.

Model selection techniques exist which take into account the diverging purposes of classification and density estimation. To understand them a simplified example using a Gaussian mixture model (GMM) [360] will be used. The reason for using this model as an example is that GMMs share essential features with HMMs, while dispensing with the complicating factor of dynamics, and are easy to visualise.

A GMM groups observations as though they arose from draws from a Gaussian distribution. This is demonstrated in figure 5.1 which shows $N = 87$ random draws (coloured discs) from three normal distributions (black dashed lines). In panel (a) a two component Gaussian mixture model (GMM) has been estimated. To do this, the data $\{s_1, \ldots, s_N\}$ were modelled as arising from the weighted sum of two normal distributions: $s \sim \pi_1 \mathcal{N}(\mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(\mu_2, \sigma_2^2)$. The values of the parameters $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \pi_1, \pi_2)$ were chosen to maximise the log likelihood $\log \mathcal{L}(\theta|\{s_i\}) = \mathbb{P}(\{s_i\}|\theta)$. The coloured lines show the estimated normal distributions and the observations have been classified

and coloured as arising from either one of the normal distributions. To classify the observations, the posterior probability of an observation, s_i, arising from component j (the distribution with parameters μ_j, σ_j), was calculated: $M_{ij} = \mathbb{P}(s_i \in \text{component } j | s_i)$ for each value of j. The observations were assigned to the component with the highest value of M_{ij}. This is known as the maximum a posteriori (MAP) assignment. In panel (c) the same model is shown but the total probability density (the sum of the probability density functions) is shown along with the log likelihood (LL). The same thing is repeated for a three component model in panels (b) and (d). So panels (a) and (b) reflect on GMMs as a method of classification, while panels (c) and (d) reflect on GMMs as a method of density estimation.

How can the two models be evaluated? For the purposes of density estimation, the estimated density of the three component model (panel (d) blue line) captures the bimodal distribution of the cluster of observations $x \in [4,8]$ better than the two component model. This is to be expected as the data were generated from a three component mixture and the log likelihood reflects this: $LL = -164$ vs. $LL = -160$ for the two and three component models respectively. From a classification perspective the situation is reversed. For both models, the cluster of data around $x \simeq 1$ is unambiguously classified as belonging to a single component. For the two component model this is also true of the observations $x \in [4,8]$. However, for the three component model, the distributions of the second and third components overlap in the small region around $x \simeq 5.8$. This means the posterior probabilities for belonging to either component, $M_{i,2}$, $M_{i,3}$, for the observations in this region will be similar. Therefore, it is not possible to unambiguously assign observations to either component $j = 2$ or $j = 3$. This is reflected in the log *classification* likelihood, LL_c which for the three component model is smaller than the two component model: $LL_c = -170$ vs. $LL_c = -165$. The log classification likelihood and the log likelihood are related by $LL = LL_c - EN$, where EN is the classification entropy [361]. The entropy of an observation, s_i, is the information entropy $H_i = \sum_j^g M_{ij} \log M_{ij}$ and measures the uncertainty with which the observation can be assigned to a given component [358]. The classification entropy is the sum of the individual entropies over the observations [190]. So although the three component model has a higher *likelihood* than the two component model, it has a lower *classification likelihood* because it cannot assign all the observations with certainty to each component.

The relationship between the Gaussian mixture model described above and a hidden Markov model is straightforward [190]. The observations in a continuous state space of the GMM $s_i \in \mathbb{R}$ map to the discrete microstate trajectories of a HMM, $s_t \in \mathbb{Z}$; the GMM components are the hidden states of the HMM h_t; the distribution parameters μ_i, σ_i are the rows of the emission matrix, $E_{i,\cdot}$; and the mixing proportions of the GMM are are stationary distribution $\boldsymbol{\pi}$ of the HMM. Considering HMMs as a coarse-graining procedure means they are aligned to the second purpose of mixture models: *classifying observations into meaningful groups*, where the 'meaningful groups' are the system's metastable states. The classification likelihood is also known as the complete-data likelihood because the classification procedure adds in a new variable, the identity of the component

associated with each observation [190]. The complete-data likelihood takes both the observation *and* the component variable into account [359]. To emphasise the difference between the likelihood and the complete-data likelihood, the former will be referred to as the observed likelihood.

As shown above, a low value of the classification entropy indicates that the emission distributions do not overlap - one of the assumptions under which HMMs are valid representations of PMMs. However, although the relative values of LL and LL_c for the two models in the above example demonstrate the difference between the classification and density estimation paradigms for mixture models, using them to assess the number of components, using for example a likelihood ratio test [249] is difficult. For a thorough discussion on the reasons for this see section 6.4 of reference [190] but briefly it arises from the fact that one can always estimate a model where the stationary distribution of one hidden states is zero thus making a smaller model (with fewer hidden states) potentially indistinguishable from the larger model. The Bayes factor approach is to integrate out all potential values of θ from the likelihood as described above. This is also possible with classification likelihood [362] although to the best knowledge of the author of this thesis, this has not been done for reversible hidden Markov models. The two likelihood approaches for HMMs may be concisely compared as follows. The integrated observed likelihood (Bayes factor approach) measures the evidence for the HMM provided by the observed microstate trajectories, while the integrated complete-data likelihood measures the evidence for the model given by the observed microstates and the given hidden states [193, 195].

5.1.6 Model selection criteria

Although the integrated observed and complete-data likelihoods are Bayesian quantities which generally require numerical approximation [252], analytic approximations exist which extend their use to models estimated using maximum likelihood [189, 190]. The most widely used approximation to the integrated observed likelihood is the Schwarz criterion [192], which up to an arbitrary factor is the Bayesian information criterion, BIC. The BIC was derived in the context of linear models and the approximations used are not valid in the finite mixture context, however, there are other theoretical and practical reasons in favour of their use [363]. The analogue of the BIC for the integrated classification likelihood is the integrated complete-data likelihood criterion, ICL [195]. The derivation of the ICL makes use of the same approximations as the BIC and so shares its drawbacks, however, in simulation experiments (for both HMMs and mixture models in general) it has performed well [190, 193, 195].

Another type of approach to selecting the number of hidden states in maximum likelihood HMMs is via minimization of the Kullback-Liebler (KL) divergence [364]. The KL divergence measures the difference between the modelled distribution and the true distribution. Two criteria which minimize this value are the Akaike-information criterion, AIC [191], and the cross-validated log-likelihood, CVLL [193].

As will be seen in this chapter, the benefits of the information criteria, BIC, ICL, and AIC, are

that they require very little extra calculation once a maximum likelihood HMM has been estimated. This is important as the search space of different models and number of hidden states may be large, rendering a more detailed Bayesian analysis for every potential model infeasible. However, these methods have drawbacks both practical and inferential. One potentially unrealistic assumption is that model selection using the AIC and BIC (and by analogy, the ICL) requires that the model representing the true data generating process must be in the model under consideration [365]. As HMMs are by design an approximation to the true dynamics, this may be an unreasonable assumption. The reasons for this differ between the AIC and BIC, however - for the Bayesian argument for the BIC see chapter 6 of reference [366]. In addition, the BIC and ICL criteria use approximations that are only valid under certain technical regularity conditions [190]. These are the same difficulties which arise for model selection using LL and LL_c. The benefit of the CVLL is that it is conceptually simple but practically one must estimate many HMMs to evaluate the number of hidden [193]. Other criteria exist for selecting the number of hidden states, for example the Penalised Marginal Likelihood criterion, PML [367] for MLE HMMs which circumvents some of the issues alluded to for the BIC, as well as a range of Bayesian model selection techniques [252, 366], however, these are not considered here.

Previous work [194] evaluated the use of the AIC, BIC and CVLL for selecting the number of microstates in Markov state models. However, choosing MSM parameters has since been superseded by the variational approach to learning Markov process (see chapter 4). This aim of this chapter is to builds on this work and investigates the use of the BIC, ICL, AIC, and CVLL to identify the number of hidden states in a HMM used for coarse-graining a MSM. It is similar to the investigation of these criteria in reference [193] but uses data simulated from the four well Prinz potential. This is an interesting extension of typical simulation benchmarks because the dynamics of the Prinz potential (a) already approximately Markovian in the observed states [64] and (b) the dynamics does not derive from an existing HMM (unlike most simulation studies which use data derived from an HMM process). The results of this chapter will be applied to the case of coarse-graining MSMs of AADH in chapter 6. The structure of this chapter is as follows: in section 5.2 the Prinz potential and the model selection criteria will be explained; section 5.3 discusses the results and section 5.4 discusses the conclusions and limitations.

5.2 Methods

5.2.1 Prinz potential

The Prinz potential [64] is shown in figure 5.2. Panel (a) shows the four well potential, $V(x)$, in blue and the stationary distribution, $\pi(x)$, showing the four metastable states, in orange. Panel (b) show the ratio of successive eigenvalues resolvable by a MSM with $\tau = 5$. The large gap between the fourth and fifth eigenvalues implies four metastable states. Panel (c) shows the exact ratio of implied timescales. The implied timescales show a large gap between the second and third implied timescale.

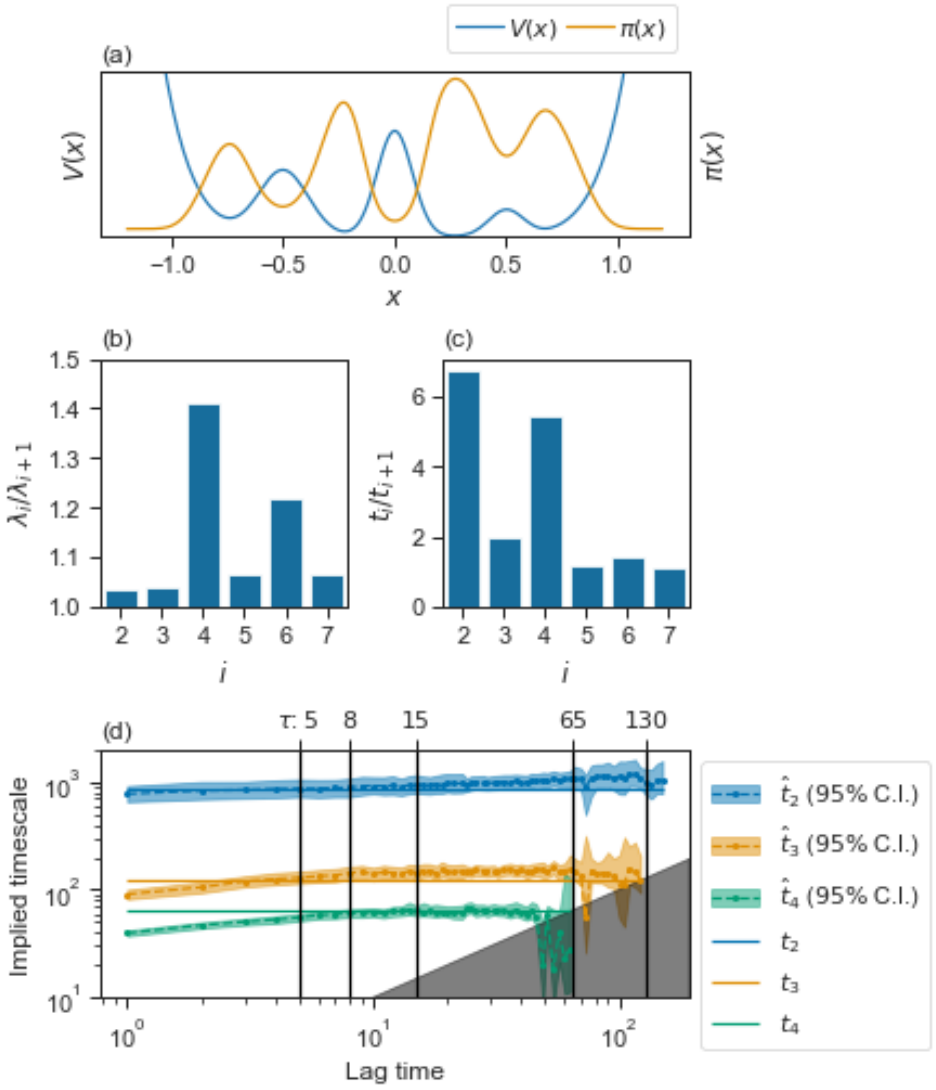


FIGURE 5.2: THE PRINZ POTENTIAL [64]. Panel (a) shows the potential $V(x)$, in blue and the stationary distribution, $\pi(x)$ in orange. Panel (b) shows the exact ratio of successive eigenvalues resolvable with a MSM with $\tau = 5$. Panel (c) shows the exact ratio of successive timescales. Panel (d) shows the estimated implied timescales, $\hat{t}_i$, as coloured dashed lines with 95 % credible intervals estimated using trajectories sampled from the potential using a Bayesian HMM with 1000 draws from the posterior. The exact values, t_i, are shown as similarly coloured solid lines. The values of $\tau = 5, 8, 15, 65, 130$ used in the model selection experiments are shown as vertical black lines.

From this potential 100 independent trajectories were sampled, initialized from random draws from the stationary distribution, discretised into 410 microstates and used as data for estimating the HMMs in this work. Each trajectory was twice the length of the longest timescale. The number of microstates was chosen as the square root of the number of observations, inline with the heuristic in reference [246]. See appendix 10 for full details of the Prinz potential and simulation details. Panel (d) shows the mean implied timescales and 95 % credible intervals as a function of the Markov lag time estimated with a Bayesian HMM. The exact timescales are also shown as solid lines. The estimated HMMs capture the exact times to within statistical uncertainty for all values of τ except for t_4 for $\tau < 8$. The number of hidden states used in these HMMs was determined by the lag time and the exact timescales of the full Prinz transfer operator (see table 10.1). For example for $\tau = 130$ only $t_2 = 844$ is resolvable so a two hidden state HMM was used. The time and rate units used throughout this chapter are in terms of the time-step used to integrate the equations of motion, $\Delta t = 0.001$, and the distance units are arbitrary, see appendix 10.

5.2.2 Model selection criteria

In the following sections, the likelihood, $\mathscr{L}$, of the HMM parameters will feature heavily and so is repeated here for convenience [169]:

$$\begin{aligned} \mathscr{L}\left(\tilde{\mathbf{T}}, \mathbf{E} | \{s_t\}\right) &= \mathbb{P}\left(\{s_t\} | \tilde{\mathbf{T}}, \mathbf{E}\right) \\ &= \sum_{\substack{\{h_t\} \in \\ \text{all paths}}} \tilde{\pi}_{h_0} E_{h_0, s_0} \prod_{t=1}^{t_{max}} \tilde{T}_{h_{t-1}, h_t} E_{h_t, s_t} \end{aligned} \tag{5.4}$$

This is the likelihood of the parameters of the transition matrix and emission matrix ($\tilde{\mathbf{T}}, \mathbf{E}$ respectively) given the trajectory of observed states $\{s_t\}$. The multiplicand represents the probability of observing Markovian transition between hidden states (the $\tilde{T}_{h_{t-1}, h_t}$ term) and then observing the observed states (the E_{h_t, s_t} term) [80, 171]. The summand represents summing the probability over all possible combinations (paths) of hidden states, while $\tilde{\pi}_{h_0} E_{h_0, s_0}$ is the probability of the initial hidden state/observed state pair [80, 171]. This summation is infeasible for even small numbers of hidden states and trajectory lengths (e.g. for 2 hidden states and a trajectory of 100 frames, there are approximately 10^{30} potential paths). The Baum-Welch algorithm [171] was developed to maximize the likelihood through expectation maximisation. An outline of the Baum-Welch algorithm can be found in algorithm 2, while the full details for maximum likelihood HMMs can be found in reference [169]. The maximum likelihood estimates of the parameters will be denoted $\hat{\theta}$, so the maximum of the observed likelihood function will be denoted $\mathscr{L}\left(\hat{\theta} | \{s_t\}\right)$.

Selecting the number of hidden states using CVLL and the AIC both minimize the *Kullback-Leibler* divergence, $D_{\mathrm{KL}}(p||q)$ [190]. This is a measure of the difference between a given probability distribution, $p(s)$ and a reference distribution, $q(s)$ [358, 364]:

$$D_{\mathrm{KL}}\left(p\,||\,q\right) = \int q(s)\log\left(\frac{q(s)}{p(s)}\right)\mathrm{d}s \\ = \int q(s)\log\left(q(s)\right)\mathrm{d}s - \int q(s)\log\left(p(s)\right)\mathrm{d}s \tag{5.5}$$

When the two distributions are the same $D_{\mathrm{KL}} = 0$. The first term is the average information of $q(s)$, also known as the information entropy [358]. This is continuous analogue of the information entropy discussed in the introduction, albeit for a different distribution. The second term is average information of $p(s)$ but averaged over reference distribution [358]. In the context of model selection, $p(s)$, is taken to be the modelled distribution $\mathbb{P}(\{s_t\}|\hat{\theta})$ and $q(s)$ is the unknown true distribution [190]. As only the latter term is dependent on the modelling choices and $D_{\mathrm{KL}} \geq 0$ (Jensen's inequality [358]) maximizing this term will lead to the model closest to the true distribution [190].

The AIC approximates the second term in equation 5.5 and is defined as [191]:

$$\mathrm{AIC} = -2\log\left(\mathscr{L}\left(\hat{\theta}|\{s_t\}\right)\right) + 2d \tag{5.6}$$

where d is the number of degrees of freedom of the model. For a reversible Markov transition matrix with g states this is: $d = {}^1\!/_2 g(g-1) + (g-1)$ [248]. The emission distribution adds $g(n-1)$, as for every hidden state g there are n probabilities which must sum to 1 [169, 171] giving $n-1$ degrees of freedom per hidden state. So the total degrees of freedom for a reversible HMM is:

$$d = {}^1\!/_2 g(g-1) + (g-1) + g(n-1). \tag{5.7}$$

The derivation of the AIC starts by approximating the true distribution, $q(s)$, with the distribution over s estimated from the data, which gives rise to the $\log\mathscr{L}\left(\hat{\theta}|\{s_t\}\right)$ term [190]. This will naturally over-fit to the data and the d term attempts to account for this. d is only equal to the degrees of freedom of the model under the assumption that the true model is under consideration in the model selection procedure [365]. The factor of -2 is there to make an equivalence with Mallows C_p [150] although this does not affect the final results. The selected model is the one which has the smallest AIC [150].

Instead of approximating $q(s)$ with $\mathscr{L}\left(\hat{\theta}|\{s_t\}\right)$ and making a bias correction, cross-validation can be used to approximate the second term of equation 5.5 [193]. In this work, the CVLL was calculated in the following way (note method was derived by the author of this thesis as the procedure in [193] deals with a different type of CV procedure):

1. The observed trajectories were split into $N = 10$ training $\{s_t\}^i$ and test $\{s_t\}^{-i}$, $i = 1, ..., N$ sets using 50:50 shuffle-split (algorithm 1).
2. For each i, an HMM was estimated using the training data $\{s_t\}^i$.
3. Calculate the log-likelihood of the training parameters using the test data, $\log\left(\mathscr{L}\left(\hat{\theta}^i\middle|\{s_t\}^{-i}\right)\right)$, with the forward part of the Baum-Welch algorithm (algorithm 2). This was achieved by taking the parameters calculated in step 2, $\hat{\theta}^i$, then using the `forward` function from the

Python package BHMM (version 0.6.3) using these parameters and the test data $\{s_t\}^{-i}$ as arguments.

4. The CVLL is the average over the splits:

$$\mathrm{CVLL} = \frac{1}{N}\sum_{i}^{N} \log\left(\mathscr{L}\left(\hat{\theta}^i \middle| \{s\}^{-i}\right)\right) \tag{5.8}$$

There are two potential points of failure in this procedure. First, the HMM may fail to converge on a given fold. Second, the 'forward' part of the Baum-Welch algorithm may fail to give a finite estimate for the log-likelihood. If either of these failures occurred, the CVLL value for that number of hidden states was considered invalid.

The BIC comes from consideration of the integrated observed likelihood, $\mathbb{P}(\{s_t\})$ used in the definition of the Bayes factor [189]:

$$\mathbb{P}(\{s_t\}|M) = \int \mathbb{P}(\{s_t\}|\theta)\,\mathbb{P}(\theta|M)\,\mathrm{d}\theta, \tag{5.9}$$

where $\mathbb{P}(\theta|M)$ is the prior distribution over the HMM parameters for a given model specification. The integrated likelihood selects the model with the greatest evidence for the observed states, i.e., the model with the highest posterior probability, given the observed states and taking into account the increased flexibility of more complex models [189, 358]. The BIC is an approximation to the logarithm of equation 5.9 and is given by [192]:

$$\mathrm{BIC} = -2\log\left(\mathscr{L}\left(\hat{\theta}\middle|\{s_t\}\right)\right) + d\log(N_{\mathrm{obs}}) \tag{5.10}$$

where d is the degrees of freedom and N_{obs} is the number of observations. The difference in BIC between two models, $\mathrm{BIC}_1 - \mathrm{BIC}_2$ is an approximation to the log of the Bayes factor, the selected model is then the one with the smallest BIC [150]. The derivation of the BIC proceeds by expanding the log of the integrand in equation 5.9, $\log(\mathbb{P}(\{s_t\}|\theta))$ in a Taylor series about $\hat{\theta}$ up to second order [150, 190]. The regularity conditions alluded to in the introduction amount to the ability to safely ignore the higher order terms in this expansion [190].

The derivation of the ICL follows an analogous path to the BIC but takes as its starting point the integrated complete-data likelihood [195]:

$$\mathbb{P}(\{(s_t, h_t)\}|M) = \int \mathbb{P}(\{(s_t, h_t)\}|\theta)\,\mathbb{P}(\theta|M)\,\mathrm{d}\theta \tag{5.11}$$

The integrated complete likelihood selects the model with the greatest evidence for the observed states *and* the hidden states [195]. As the hidden states are not observed they are taken to be MAP assigned values: $h_t = \mathrm{argmax}_j\, M_{s_t,j}$. The ICL is an approximation to $\log\left(\mathbb{P}(\{(s_t, \hat{h}_t)\}\middle|M\right)$ and is given by [195]:

$$\begin{aligned}\mathrm{ICL} &= -2\log\left(L\left(\hat{\theta}\middle|\{s_t\}\right)\right) + d\log(N_{\mathrm{obs}}) + 2\cdot EN(\mathbf{M}) \\ &= \mathrm{BIC} + 2\cdot EN(\mathbf{M})\end{aligned} \tag{5.12}$$

The EN term is classification entropy given by [195]:

$$\begin{aligned} EN(\mathbf{M}) &= \sum_{t}^{N_\mathrm{T}}(-1)\sum_{j}^{g} M_{s_t,j}\log\left(M_{s_t,j}\right) \\ &= \sum_{t}^{N_\mathrm{T}} H(s_t;\mathbf{M}) \end{aligned} \tag{5.13}$$

Here $\mathbf{M}$ is the membership matrix $M_{ij} = \mathbb{P}(h = j|s = i)$ and H is the information entropy $H(s_t;\mathbf{M}) = -\sum_j M_{s_t,j}\log\left(M_{s_t,j}\right)$. This entropy quantifies the uncertainty with which the model assigns the given observed state to a hidden state [358]. For example in a two hidden state system, given an observed state which could belong in hidden state 1 with probability 50 % or in hidden state 2 with probability 50 %, then the entropy for that observation is:

$$\begin{aligned} H(s;\mathbf{M}) &= -\sum_{j} M_{s_t,j}\log\left(M_{s_t,j}\right) \\ &= -{}^1\!/_2\log({}^1\!/_2) - {}^1\!/_2\log({}^1\!/_2) \\ &= \log(2) \end{aligned} \tag{5.14}$$

5.2.3 Criteria calculation details

There are a number of practical details in calculating the information criteria which need to be addressed.

The number of observations, N_obs, needed for the BIC and ICL, was calculated as the number of pairs of observations which go into the count matrix. Using the sliding window count method this is:

$$N_\mathrm{obs} = N_\mathrm{traj}\cdot\frac{(N_\mathrm{T}-\tau)}{\Delta t} \tag{5.15}$$

where N_traj is the number of trajectories, N_T is the length of each trajectory, τ is the Markov lag time and Δt is the trajectory time-step. The total number of frames is $N_\mathrm{frames} = N_\mathrm{traj}\cdot N_\mathrm{T}/\Delta t$

The classification entropy was calculated using the hidden state probabilities calculated in the final iteration of the Baum-Welch (B-W) algorithm (algorithm 2):

1. For each observed state in a trajectory, s_t, the conditional probability, $\gamma_i(t) = \mathbb{P}(h_t = i \mid s_t, \theta)$, was extracted from the final iteration of 'update' part of the B-W algorithm.
2. The entropy was calculated for each frame of the trajectory, $H(t) = -\sum_i \gamma_i(t)\log\left(\gamma_i(t)\right)$ and then summed over the N_T frames of a trajectory and then over the N_traj trajectories: $EN = \sum^{N_\mathrm{traj}}\sum_t^{N_\mathrm{T}} H(t)$.
3. N was scaled by a factor of $N_\mathrm{obs}/N_\mathrm{frames}$ to account for the fact that an observation is a pair of states $(s_t, s_{t+\tau})$.

A second method was available, which in principle should give the same answer but which in practice diverged by up to a factor of three from the above method. The entropy was calculated using the

membership matrix, itself calculated from the emission matrix, $\mathbf{E}$, and the stationary distributions of the hidden $\widetilde{\boldsymbol{\pi}}$ and observed $\boldsymbol{\pi}$ states:

$$M_{ji} = E_{ij}\frac{\widetilde{\pi}_i}{\pi_j} \tag{5.16}$$

where i labels the g hidden states and j labels the n observed states. The entropy was calculated using $\mathbf{M}$ as:

$$\mathrm{EN}(\mathbf{M}) = N_{\mathrm{obs}}\sum_{j}^{n}(-1)\pi_j\sum_{i}^{g}M_{ji}\log\left(M_{ji}\right) \tag{5.17}$$

The reason for the difference was due to error accumulated in the values of $\mathbf{M}$ from equation 5.16. This was due to noise in the poorly sampled observed states, see section 10.2 and figure 10.1. However, in situations when the largest reversible connected set of hidden states is smaller than total number of hidden states, the values of γ would need re-normalizing. In this case (in particular in chapter 6) the second method will be more convenient.

All HMM and MSM fitting was performed in Python 3.7 using the packages PyEMMA (version 2.5) [253], NumPy (version 1.19) [338], Pandas (version 0.23) [339], Matplotlib (version 3.3) [340], Seaborn (version 0.10) [341] and the Jupyter Project (version 4.6) [342]. The calculation of the cross-validated log-likelihood used the Python package Bayesian hidden Markov model toolkit (BHMM, version 0.6.3). All scripts to calculate the AIC, BIC and ICL were written by the author of this thesis.

5.2.4 Model selection

The model selection criteria were used to select the optimum number of hidden states in a maximum likelihood HMM, using the discrete trajectories sampled from the Prinz potential. Five different values of the Markov lag time were used: $\tau = 5,\ 8,\ 15,\ 65,\ 130$. These values were chosen because they resolve, respectively $7,\ 5,\ 3,\ 2,\ 1$ implied timescales in the full MSM state space and as the top three of these timescales are dominant (figure 5.2 panel (b)), these values of τ resolve $4,\ 4,\ 4,\ 3,\ 2$ metastable states respectively [169].

For each value of τ maximum likelihood HMMs were estimated with $g = 2-10$ hidden states. For each of the 45 model specifications (different values of τ and g) the model selection criteria were calculated and the number of hidden states selected by each was compared to the true value.

5.3 Results and discussion

The selected number of hidden states using each criterion are shown in table 5.2 where the asterisk denotes when a criterion selects the correct number of hidden states. The relative values of the selection criteria are shown in figure 5.3. Each row, (a) - (e), corresponds to models estimated with a different value of the Markov lag time $\tau = 5,\ 8,\ 15,\ 65,\ 130$. Each column, (i) - (iv), corresponds

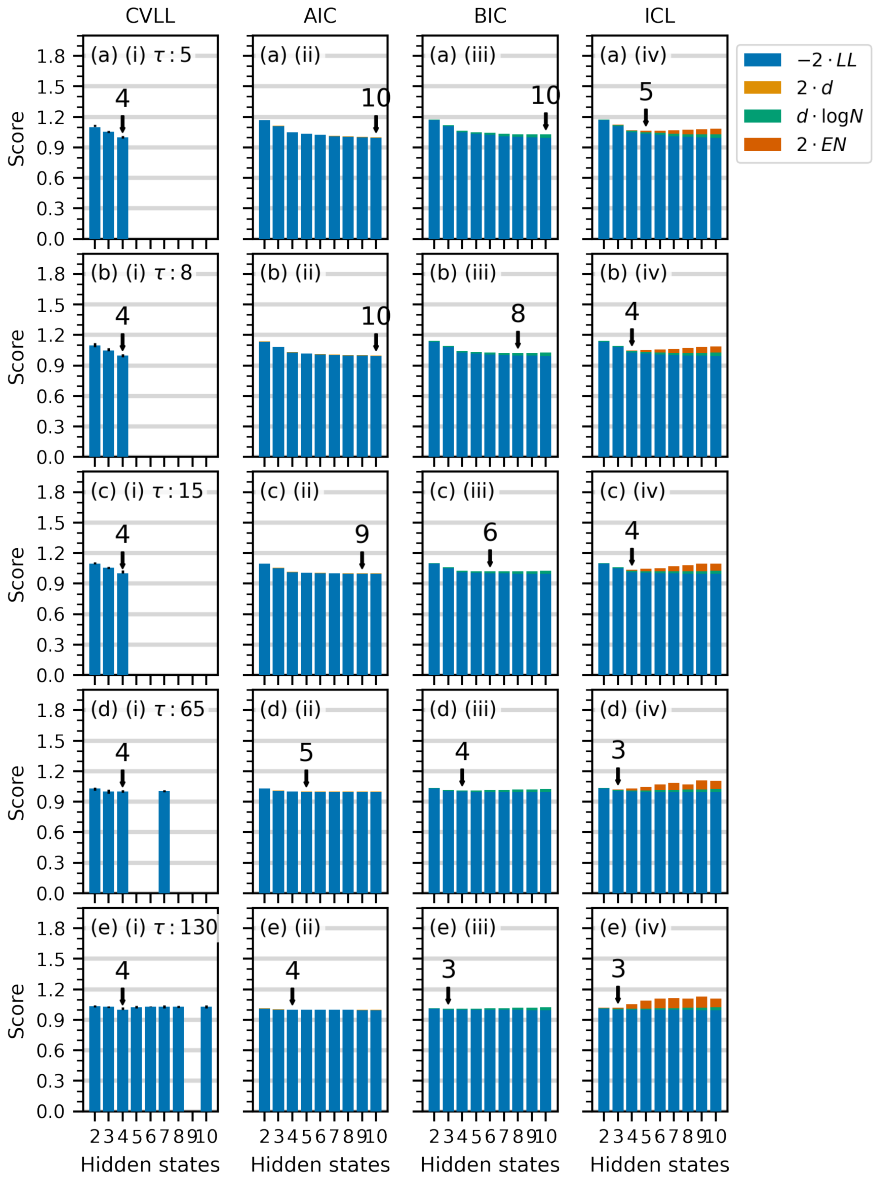


FIGURE 5.3: HIDDEN STATE SELECTION CRITERIA. Rows (a) - (e) show the selection criteria for HMMs with $\tau = 5, 8, 15, 65, 130$ respectively. The best performing number of hidden states is indicated by an arrow. Column (i) shows the CVLL. Column (ii) shows the AIC. The log-likelihood term is shown in blue and the degrees of freedom penalty ($2d$) is shown in orange. Column (iii) shows the BIC. The penalty term $d \cdot \log N_{obs}$ is shown in green. Column (iv) shows ICL. The classification entropy penalty term $2 \cdot EN$ is shown in red. Missing values indicate the failure of the HMM to converge. All values have been scaled so the minimum value in each panel is 1.

τ	g^{true}	CVLL	AIC	BIC	ICL
5	4	$4^*(0)$	10(6)	10(6)	5(1)
8	4	$4^*(0)$	10(6)	8(4)	$4^*(0)$
15	4	$4^*(0)$	9(5)	6(2)	$4^*(0)$
65	3	4(1)	5(2)	4(1)	$3^*(0)$
130	2	4(2)	4(2)	3(1)	3(1)

TABLE 5.2: HIDDEN STATE SELECTION RESULTS. The selected number of hidden states, $\hat{g}$, by the CVLL, AIC, BIC and ICL for each value of τ. The true values, g^{true} are also shown. The asterisk highlights where $\hat{g} = g^{\text{true}}$. The number in parentheses shows $\hat{g} - g^{\text{true}}$.

to the different model selection criteria, CVLL, AIC, BIC, and ICL. The minimum value of each criterion for each model is highlighted with an arrow indicating the selected number of hidden states, $\hat{g}$. The values are scaled so the value at the selected number of states the value of the criterion is equal to 1. The coloured bars show the contributions of the different parts of each score. The blue bars shows the log-likelihood terms of equations 5.6, 5.10 and 5.12 i.e. $-2 \times \log\left(\mathscr{L}\left(\hat{\theta}\middle|\{s_t\}\right)\right)$. In the case of CVLL, the blue bars are the cross-validated equivalent. The various penalty terms are shown in orange ($2d$, the AIC penalty), green ($d \log N$, the BIC penalty) and red ($2 \cdot EN$, the classification entropy).

The ICL performs best by correctly identifying the number of hidden states for $\tau = 8,\ 15,\ 65$. It fails at $\tau = 5$ where the hidden state dynamics are not quite Markovian (figure 5.2 panel (d)). Although the selected value of 5 is close to the true value of 4, the ICL does not discriminate between $g = 4 - 7$: their ICL values vary by less than 1 %, as shown in figure 5.3 panel (a)(iv). The ICL also fails at $\tau = 130$, however, the minimum value of the ICL is similar to the value for the true number of hidden states, $g = 2$, and is significantly different to the values for $g \geq 4$, as shown in panel (e)(iv). In this case the ICL does distinguish between two sets of values of g, which include the true value on the one-hand, and the remaining values on the other. This behaviour is in contrast to the results in reference [193] in which the ICL correctly identified the number of hidden states for well separated emission distributions, and with large numbers of observations for less well separated distributions. However, for smaller numbers of observations and less well separated clusters, the ICL *under-estimated* the number of hidden states. The ICL also under-estimated the number of clusters in the finite mixture context when the clusters are not well separated [195].

The CVLL correctly selects four states for $\tau = 5,\ 8,\ 15$, however, this was due failure of the cross-validation to produce a finite answer on some of the cross-validation folds. For example, for $\tau = 5$, at least one cross-validation fold did not estimate the out-of-sample log-likelihood for $g \geq 4$. The remaining values are shown in figure 5.3 column (i). This causes problems with interpretation as is it is not clear whether failure is due to the inefficiency of the cross-validation procedure or whether the given number of hidden states really has zero out-of-sample likelihood. The former is more likely given that models estimated on 100 % of the data do converge and give interpretable

answers. Given the lack of convergence for many of the values of g comparison with the literature is difficult. The results in reference [193] show that the CVLL behaves similarly to the ICL but with less discrimination between values of g i.e., in repeated experiments the distribution of selected values of g was wider for the CVLL. In contrast, the results for $\tau = 130$ in figure 5.3 panel (e)(i) show the CVLL over-estimates the number of hidden states. Given the poor performance of the CVLL in this experiment it will not be discussed further here.

The AIC overestimates for every value of τ and as τ increases the values of the AIC discriminate less between each value of g: for $\tau = 130$ the AIC for all g are within 2 % of each other. This is in contrast to the results in reference [193] in which the AIC selected the correct or underestimated the value of g. However, in simulation studies for finite mixtures (without the Markovian dynamics of the hidden states) the AIC frequently over-estimated the number components [368, 369]. The BIC also overestimates the number of hidden states for all values of τ but only by 1 for $\tau = 65$ and 130. This is in contrast to the results in reference [193] for which the BIC behaved similarly to the ICL and either estimated correctly or under-estimated the number of hidden states. In addition, for finite mixtures the BIC has also been shown to under-estimate the number of components [195].

Although the AIC, BIC and ICL are derived from different starting points, they all take the form of the log-likelihood plus a penalty term, b:

$$-2\log\left(\mathscr{L}\left(\hat{\theta}\middle|\{s_t\}\right)\right) + b \tag{5.18}$$

The b term in each case penalises the complexity of each model. The behaviour of these criteria can be understood in terms of the interplay between the likelihood and penalty terms. The log-likelihood, the blue bars in columns (ii)-(iv) of figure 5.3, monotonically increases with g for all values of τ (this is shown as a decrease due to the -2 in the definition of the criteria). This is most pronounced for small values of τ (compare panel (a)(ii) to (e)(ii)), and demonstrates both over-fitting and the HMMs ability to capture the fast relaxation processes of the Prinz dynamics. Consider the $g = 10$ model selected by the AIC for $\tau = 8$, figure 5.4 (the results are similar for $\tau = 5$). This figure shows the sign structure of the exact relaxation processes (panels (a) - (c)) and those estimated from the HMM (panels (d) - (f)). The HMM captures the sign structure of the second and fifth relaxation process, panels (d) and (e), as they have associated timescales larger than τ, ($t_2 = 844.4, t_5 = 11.9 > \tau = 8$) and are thus resolvable. The 10th estimated relaxation process (panel (f)) only approximate the true relaxation process (panel (c)) as the estimated timescale $\hat{t}_{10} = 4.0$ is less than the lag time, $\tau = 8$. For larger values of τ and $\tau = 130$ in particular (figure 5.3 panel (e)(ii)), the likelihood remains constant. This is because many of the estimated HMM relaxation processes over-fit to noisy fluctuations in the data.

For the AIC the penalty term, $b = 2d$ (orange bars in figure 5.3, column (ii)), is there to correct the approximation of the KL divergence by log-likelihood. It increases proportional to g^2 (equation 5.7) but only affects the selected g for $\tau = 15,\ 65,\ 130$ (figure 5.3 panels (c)(ii), (d)(ii) and (e)(ii)). The origin of the BIC penalty term, $b = d\log N_{obs}$ (green bars, column (iii)) is to correct the approximation of the integrated likelihood by the maximum log-likelihood. The BIC over-estimates

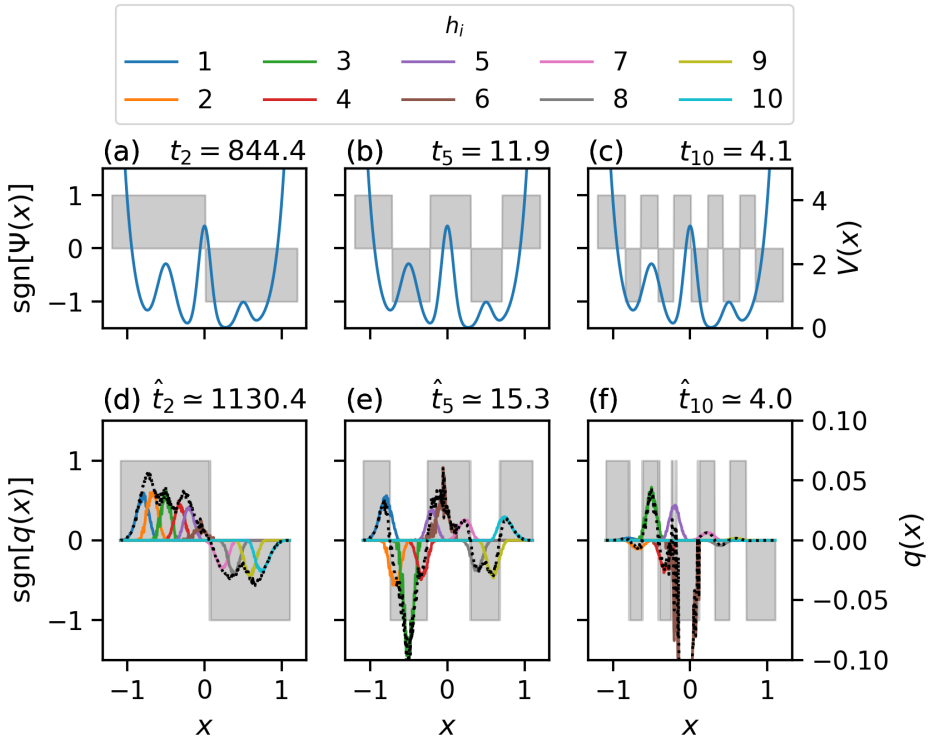


FIGURE 5.4: COMPARISON OF ESTIMATED AND TRUE PRINZ POTENTIAL DYNAMICS. The true Prinz potential are compared with a HMM with $\tau = 8$ and $g = 10$ hidden states. Panels (a) - (c) shows the sign structure of the 2nd, 5th and 10th right eigenvector of the Prinz potential ($\mathrm{sgn}[\Psi(x)]$, shaded area). The Prinz potential ($V(x)$, blue solid line) is shown for reference. The exact timescales are labelled on the top right as $t_{2/5/10}$. Panels (d) - (f) show the sign structure of the hidden state relaxation processes, projected onto the observed states. The eigenvectors projected onto the observed state basis, $q_{2/5/10}(x) = \sum_i E_{i,x} \cdot \tilde{\Psi}_{2/5/10}(i)$, are shown as dotted lines, the summands are shown as coloured lines. The shaded areas are $\mathrm{sgn}[q_{2/5/10}(x)]$. The estimated timescales are labelled on the top right as $\hat{t}_{2/5/10}$.

the number of hidden states albeit by a smaller number than the AIC, due to the penalty term rising faster with g by a factor $\log N_{\mathrm{obs}}/2$. As pointed out in reference [194] when using the sliding window method for calculating the count matrix the value of N_{obs} will be overestimated. However, as the dynamics is approximately Markovian for $\tau > 5$ the difference between the sliding window and sample count methods will be negligible.

The ICL penalty term $b = d \log N + 2 \cdot EN$ corrects the approximation to the integrated complete-data likelihood by the log-likelihood. It is comprised of the BIC penalty term (green bars in figure 5.3) and the entropy term (red bars). This is associated with increasing g directly through the BIC penalty term, $\simeq g^2 \log N_{obs}$, and indirectly due the overlap of emission distributions. This is demonstrated in figure 5.5. Panel (a) shows the emission distribution for a two state HMM. These two distributions only overlap around $x = 0$. Panel (d) shows the information entropy at each value of

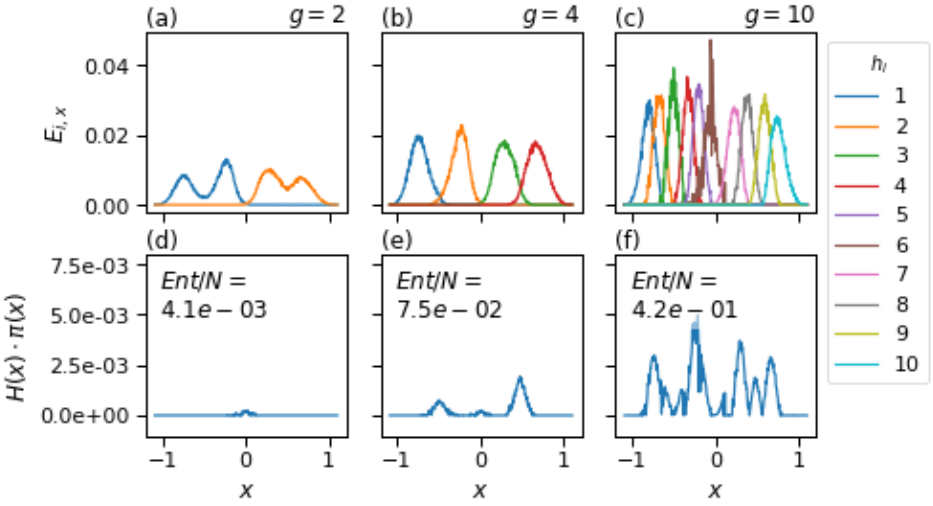


FIGURE 5.5: THE CLASSIFICATION ENTROPY OF HMMS. Panels (a)-(c) show the emission distributions of HMMs with $\tau = 8$ and $g = 2, 4, 10$ respectively. Each coloured line represents the emission distribution, $E_{i,x}$, of the hidden states, i. Panels (d) - (f) show the information entropy for observed state at x, weighted by the stationary distribution over the observed states: $\pi(x)H(x)$. The label shows the average classification entropy per observation $EN_{\text{ave}} = \sum_x \pi(x)H(x)$.

x, weighted by the stationary distribution over the observed states, $\pi(x)$[1]. The information entropy is zero almost everywhere as each observed state can be assigned unambiguously to a hidden state. The exception is around $x = 0$ where the entropy reaches its highest possible value of $\log 2$. However, the average classification entropy per observation, $\sum_x \pi(x)H(x)$ is low as the fraction of observations at $x = 0$, $\pi(0)$, is negligible. As the number of hidden states increases, panels (b) and (c), the entropy increases because the emission distributions overlap more, and the average entropy increases because they overlap in regions which are visited more often i.e. where $\pi(x)$ has significant density (panels (e) and (f)). As column (iv) of figure 5.3 shows, the entropy penalty is the source of the success of the ICL in selecting the correct number of hidden states.

Minimizing the entropy penalty alone is similar to maximizing the crispness/scaling condition in PCCA+ (equation 4.19 in reference [159]) in that it maximizes the number of observed states that are unambiguously assigned to one hidden state. However, the minimum entropy solution by itself (ignoring the other terms) will always favour two hidden states separated by the slowest relaxation process, which for small τ does not capture the potential other metastable states. The ICL balances the need for a 'crisp' assignment with the need for hidden states needed to accurately model the transition matrix.

[1] the information entropy here the same calculation which determined the transparency of the observations in figure 5.1 panel (b).

5.4 Conclusions

Four model selection criteria have been compared for identifying the number of hidden states in HMMs of dynamics simulated from the four-well Prinz potential. The four criteria fall into two categories - those that aim to minimize the Kullback-Liebler divergence, the CVLL and the AIC, and the those that maximize a type of integrated likelihood, the BIC and ICL. The CVLL was of limited usefulness because it was unable to produce results for a significant proportion of the models tested and because of its relatively large computational requirement. The AIC and BIC both overestimated the number of hidden states although the BIC by fewer states than the AIC. These results do not match the results from previous studies on selecting the number of components in mixture models and in HMMs which tend to underestimate the number of hidden states. The ICL, which maximizes the integrated complete-data likelihood performed best by correctly identifying three out of five hidden states and where it failed it only overestimated by one extra hidden state. The main limitations of this work is that it did produce a statistical estimate of the selected number of hidden states. In other studies [195] the criteria are judged on an ensemble of models with similar characteristics and also on a range models with different characteristics.

The data from the four state model used to test these criteria limits the conclusions that can be drawn. The dynamics used to generate the trajectories was Markovian by construction meaning observations were independent of one another. This ensured that the underlying assumption of independent and identically distributed (i.i.d.) observations, which is required of all criteria tested here, was met. However, in molecular dynamics simulations more commonly use for biomolecular simulations, frames are correlated at short times which violates these assumptions. Separate testing with molecular dynamics simulations is required to see how this lack of independence affects these criteria. The sampling of the data was also unrealistic in other ways: first, the trajectories were sampled from equilibrium distribution; second, the total length of simulation data was 200 times the longest relaxation timescale. Neither of these conditions is met in typical simulations. A second limitation is the nature of the Prinz potential which is smooth, with none of the hierarchical structure arising from the many degrees of freedom typical of biochemical systems. One potential remedy for this deficiency is to introduce some randomness to the potential to simulate a more rugged potential energy landscape and re-test these criteria. The effect of a more complex energy landscape would be to increase the number of potential energy minima and dynamical processes. It is unclear exactly how the BIC, AIC and ICL would perform under these more realistic conditions. The likelihood function would likely increase more rapidly with the number of hidden states as the model tries to fit to more noisy fluctutions (in the case of a fewer, shorter trajectories) or to more complex free energy surface. This would increase the probability of models being selected with too many hidden states. However, it is not clear how the data dependent penalties of the BIC and ICL (i.e., the number of observations and the classification entropy) would compensate for this increase and requires further testing.

Potential avenues for testing the ICL further with model potentials would be to both increase

the "roughness" of the four-well potential and to decrease the amount of data used to estimate the HMMs. This would allow a more realistic assessment of its applicability while still having a definitive number of metastable states, which is not always easy to ensure with real systems.

However, under the idealised conditions of the model the ICL is a promising candidate for HMM state selection. The integrated complete-data likelihood is a natural criterion for the purpose of coarse-graining MSMs of conformational dynamics. The penalisation term in the ICL is aligned with assumptions that make metastable Markov processes amenable to coarse-graining with a HMM. The purpose of coarse-graining is to provide an interpretable model of dynamics which means balancing simplicity and accuracy. Part of the simplicity of a coarse-grained model is being able to interpret given structures (microstates) as belonging, unambiguously, to a particular metastable state. Considering the integrated classification likelihood naturally penalises less interpretable solutions by considering the models' evidence for both the observed states and the classification into hidden states.

CHAPTER 6

AROMATIC AMINE DEHYDROGENASE

Contributions

All work in this chapter was performed by the author of this thesis with the following exception. Dr Kara Ranaghan prepared the PDB file of the crystal structure of AADH in Schiff base form after reaction with tryptamine, PDB accession code 2AGY [197]. They added the missing hydrogen atoms, determined the protonation states of titratable residues, created the disulphide bridges and parameterized the tryptophan tryptophylquinone (TTQ)/tryptamine Schiff base (structure A in figure 6.2, custom residue name TTW) for use with the CHARMM-22 forcefield [370] as part of the computational part of references [197, 371, 372]. The first 23 residues of the D and H chains, which were unobserved in the crystal structure, were not modelled.

6.1 Introduction

Aromatic amine dehydrogenase (AADH) catalyses the oxidation of primary aromatic amines, such as tryptamine, to the corresponding aldehyde and ammonia [197]. Amines are natural by-products of human activity and and their degradation is an is an important part of the natural cycle which maintains their balance within organisms and the environment [373]. AADH containing soil bacteria utilise aromatic amines as a source of energy and carbon [374] and thus play an important part in maintaining this balance. The importance of AADH as an enzyme from a chemists' point of view derives in large part from its large kinetic isotope effect and its purported link to controversy surrounding the role of protein dynamics in enzyme catalysis. This link and the role that conformational dynamics can play to resolving this controversy will be explored in the this chapter.

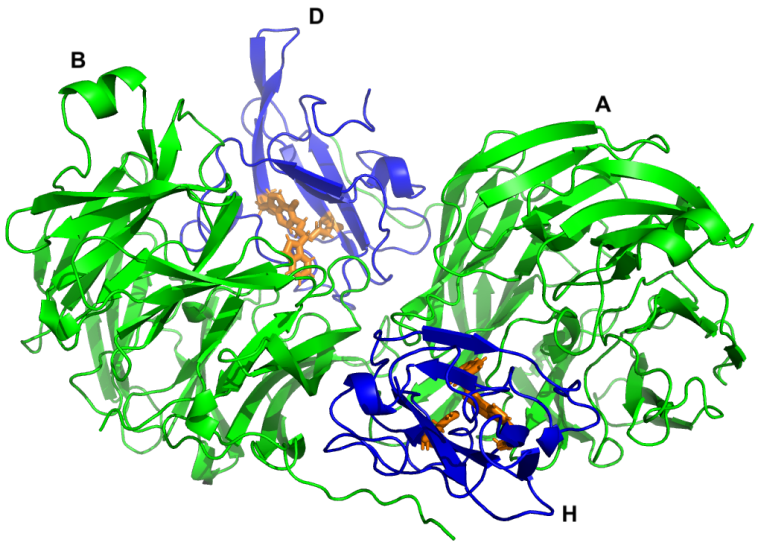


FIGURE 6.1: CRYSTAL STRUCTURE OF AADH. PDB accession code: 2AGY [197]. The α chains, A and B, are shown in blue, the β chains, D and H, are in green. The tryptophan tryptophyl quinone (TTQ) prosthetic groups after reaction with the substrate tryptamine are shown in orange.

6.1.1 Structure, reaction mechanism and kinetics

AADH was first isolated from a soil bacterium of the genus *pseudomonous* where its molecular weight and amino acid composition determined, and a catalytic chromophore identified [375] (in the references cited below however, AADH from the soil bacterium alcaligenses faecalis is used). The chromophore was later identified as a tryptophan tryptophylquinone (TTQ) group using resonance Raman spectroscopy and by comparison with methylamine dehydrogenase (MADH), which also contains a TTQ group [374]. The reaction proceeds in two stages: the first step is the reductive half-reaction where the substrate is oxidized by the TTQ group; the second stage is the oxidative half-reaction whereby two electrons are transferred to the copper containing protein Azurin. A c-type cytochrome then transfers the electrons into the respiratory chain [197].

The structures of the substrate-free enzyme and of several enzyme-substrate intermediates were solved to determine the mechanism of the first stage, the oxidation of tryptamine [197] (this mechanism will be discussed further below). In particular, the structure of the Schiff-base intermediate was solved (PDB accession code 2AGY) and is the basis for the work in this chapter. Other structures have been solved for a variety of purposes. The structures after reaction with a number different substrates were solved in order to explore the role of conformational changes required for the second stage, the transfer of electrons to the Azurin protein [376]. An AADH-azurin complex structure was solved in order to explore the electron transfer mechanism and to compare it to MADH [377]. The kinetic isotope effect was studied for a range of benzyl amine substrates and

for this type of substrate, revealed that structural rearrangements prior to hydrogen transfer are rate limiting, rather than the hydrogen transfer itself [378].

AADH has an $\alpha_2\beta_2$ structure and is shown in figure 6.1. The larger α chains have a mass of $\simeq 39\,\mathrm{kDa}$ (shown in green and labelled A and B) while the smaller β chains (shown in blue, labelled D and H) have a mass of $\simeq 18\,\mathrm{kDa}$. The D and H chains contain the TTQ prosthetic group that forms part of the active site (shown in orange, after reaction with tryptamine).

The reaction mechanism of AADH was proposed in reference [379] and more fully elucidated by a combination X-ray crystallography, experimental and QM/MM studies in references [197, 371], all using tryptamine as a substrate. The reaction mechanism is shown in figure 6.2 which has been adapted from figure 2 of reference [197]. The following description is also taken from reference [197]. The TTQ prosthetic group, attached to the β sub-unit, is shown in purple, the protonated tryptamine substrate is shown in blue and the Asp128 residue is shown in red. The mechanism starts with the enzyme substrate complex of the protonated tryptamine situated next to the TTQ group and Asp128 residue. The tryptamine is deprotonated by oxygen 1 of the Asp128 residue via a bridging water molecule (step 1). The nitrogen atom on the tryptamine attacks one of the carbonyl groups of the TTQ residue to form a carbinol-amine intermediate (step 2) that then goes on to form the iminoquinone intermediate labelled 'A' (step 3). Intermediate A was not observed directly but was inferred from the crystal structure of an analogous complex using phenylhydrazine in place of tryptamine. Step 4, shown boxed, is the rate limiting step and involves the tunnelling of a proton from the tryptamine carbon atom adjacent to the nitrogen, to an oxygen atom of the Asp128 residue. The proton is shown here accepted by oxygen 2 of the Asp128 carboxylate group, but in principle oxygen 1 could also serve as an acceptor. The oxyanion on the TTQ/substrate group (hereafter referred to as TTW) is then neutralized by protonation from Asp128 via the hydrogen atom on the protonated Schiff base (step 5). Water is introduced which in step 6 attacks the Schiff base to form a carbinolamine intermediate that is then oxidized in step 7. The carbinolamine is then hydrolysed in step 8 and releases the aldehyde.

The experimental free energy barrier for this reaction is approximately[1] $12.7\,\mathrm{kcal\,mol^{-1}}$ at $T = 300\,\mathrm{K}$. The rate limiting proton abstraction can proceed to either OD1 or OD2 of the Asp128. These two atoms are distinguished by the hydrogen bonding network: OD2 is hydrogen bonded to Trp160, and OD1 to Thr172 as shown in figure 6.7. These two pathways were first elucidated using semiemperical QM/MM methods [197]. Ab-initio QM/MM methods were later used to calculate the potential energy surface of the reaction along these two paths using local coupled cluster theory with a large basis set (LCCSD(T)/(aug)-cc-pVTZ) [372] and found barrier heights of approximately 16.8 and $14.1\,\mathrm{kcal\,mol^{-1}}$ to OD1 and OD2 respectively. Combining these barrier heights with estimates of the zero-point energy and tunneling effects, the authors of reference [372] estimated effective barriers of 10.6 and $8.3\,\mathrm{kcal\,mol^{-1}}$. With an entropic contribution to the free energy estimated to be between $0.4\,\mathrm{kcal\,mol^{-1}}$–$4\,\mathrm{kcal\,mol^{-1}}$ [380, 381], the two pathway reaction mechanism is compatible

[1] No error estimate was given in reference [197].

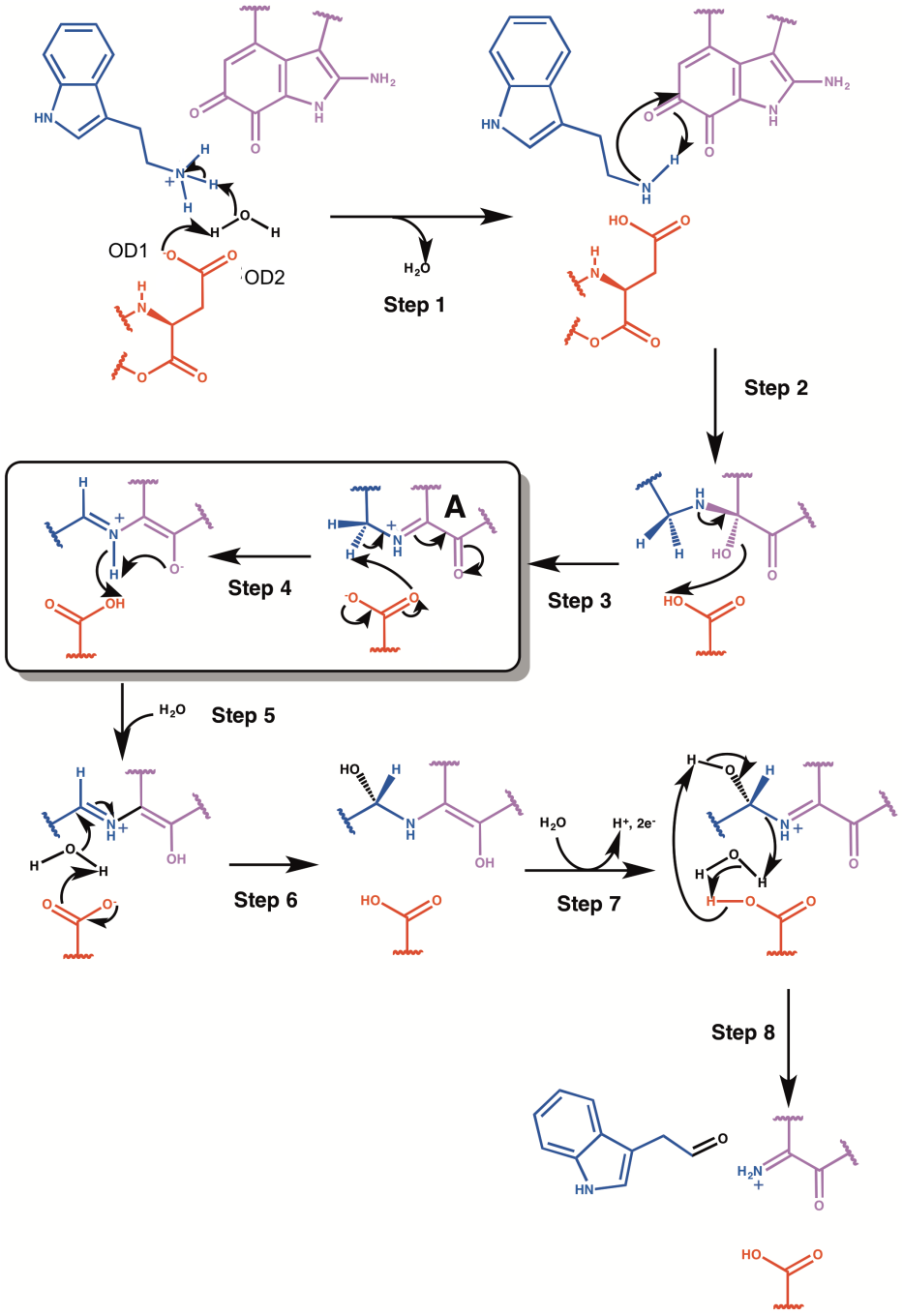


FIGURE 6.2: REACTION MECHANISM OF AADH. This reaction mechanism was proposed by reference [197]. The detailed reaction mechanism is taken from figure 2 of that reference, alterations were made to improve legibility of the arrows denoting movement of electrons. The TTQ prosthetic group is shown in purple, the Asp128β group in red and the tryptamine substrate in blue. The boxed step 4 is the rate limiting tunnelling step. In step 1, a water molecule in the active site is used to transfer a hydrogen atom from the protonated tryptamine to the deprotonated Asp128. In step 5 a water molecule is transferred from the bulk solution to form the carbinolamine intermediate after step 6. Step 7 represents the introduction of water from bulk solvent used in step 8 and two consecutive long-range electron transfers to the electron carrier Azurin.

with the observed reaction free energy of $12.7\,\mathrm{kcal\,mol^{-1}}$.

Note: in previous AADH simulation studies [197, 371, 372], the authors labelled OD1 as O2 and OD2 as O1, this convention is not adopted here as the accessible conformations in the simulations later show, both OD1 and OD2 can hydrogen bond to both Trp160 and Thr172, so the force-field atoms names will be retained.

6.1.2 Kinetic isotope effect

The primary deuterium KIE (1° KIE) is defined as the ratio of microscopic rate constants, $\frac{k_\mathrm{H}}{k_\mathrm{D}}$ for the following reactions [382]:

$$\mathrm{C{-}H} + \mathrm{A} \xrightarrow{k_\mathrm{H}} \mathrm{C} + \mathrm{A{-}H} \tag{6.1}$$

$$\mathrm{C{-}D} + \mathrm{A} \xrightarrow{k_\mathrm{D}} \mathrm{C} + \mathrm{A{-}D} \tag{6.2}$$

where the 'A' species is a generic hydrogen *acceptor*, the carbon atom is the *donor* atom, and 'H' and 'D' are hydrogen and deuterium atoms or ions. For AADH the donor is C1 (the carbon α to the substrate amine group), the acceptor is OD1 or OD2 of the carboxylate ion in Asp128, and the transferred proton comes from HI-2 shown in figure 6.7 for tryptamine substrate.

Treating the C—H bond as a harmonic oscillator with fundamental frequency $\omega_\mathrm{H} \simeq 3000\,\mathrm{cm^{-1}}$ and at low temperatures ($\frac{\omega}{k_\mathrm{B}T} \gg 1$), the semi-classical extension of transition state theory (TST) gives the following expression for the KIE [198]:

$$\ln \frac{k_\mathrm{H}}{k_\mathrm{D}} = \frac{1}{2k_\mathrm{B}T}\left(1 - \sqrt{\frac{m_\mathrm{H}}{m_\mathrm{D}}}\right)\omega_\mathrm{H} \tag{6.3}$$

where k_B is the Boltzmann constant and $m_{H/D}$ is the mass of the hydrogen and deuterium atoms. The KIE in this model arises exclusively from the difference in zero-point energy of the two bonds which results in two features:

1. at room temperature the maximum KIE will not exceed $\frac{k_\mathrm{H}}{k_\mathrm{D}} \simeq 8$, and
2. the KIE is dependent on temperature.

KIEs for AADH have been observed which are both above this semi-classical limit: 8.6–12.9 with dopamine [200, 383], 55 ± 6.0 with tryptamine [197] and independent of temperature over the range $281\,\mathrm{K}$–$294\,\mathrm{K}$ [197]. AADH is not alone in exhibiting large and temperature independent KIEs. Methylamine dehydrogenase (MADH), like AADH, also contains a TTQ co-factor, but instead catalyses aliphatic primary amines. The MADH KIE with methylamine has been reported as 17.2 [201] and 16.8 [202] and both independent of temperature. Soybean-lipoxygenase (SLO) catalyses the oxidation of linoleic acid with a weakly temperature dependent KIE of approximately 80 [203, 204]; dihydrofolate reductase (DHFR) catalyses the reduction of 7,8-dihydrofolate (H2F) to 5,6,7,8-tetrahydrofolate with a temperature independent KIE, albeit with a KIE of 2 [205, 206]; other

systems have also been investigated, see the reviews in references [207, 208]. Observed KIEs elevated above the semi-classical limit, along with observations of the temperature dependence of the KIEs, as well as the relationship between deuterium and tritium secondary KIEs (where the substituted hydrogen is α to the reacting bond) provided evidence that tunneling was occurring, see reference [199] for a history of the evidence.

6.1.3 Reaction rate theories incorporating tunneling

A number of theories have been developed to explain enzymatic reaction rates incorporating tunneling. The two central theories are: corrections to the semi-classical TST theory (from which equation 6.3 was derived) [384] and full-tunneling models [198, 204, 214, 385]. Tunneling is a quantum mechanical processes whereby a particle (in the case of AADH, a proton) can be transferred through a potential energy barrier without the required amount of kinetic energy [215, chapter 3, 5]. Alternatively [386], the de-Broglie wavelenth of a proton at biological temperatures is approximately $1\,\text{Å}$, meaning uncertainty in its position is of the order of distance travelled during the course of the reaction, implying that quantum mechanical effects will be important.

The tunneling correction to semi-classical TST amounts to a pre-factor in the expression for the rate [209]:

$$k(T) = \gamma(T)\left(\frac{k_\text{B}T}{h}\right)\left(C^0\right)^{1-n}\exp\left[\frac{-\Delta G^{\text{TS},0}(T)}{RT}\right] \tag{6.4}$$

where $G^{\text{TS},0}(T)$ is the activation free energy (including the effect of zero point energy), RT the average thermal energy per mole, $(k_\text{B}T/h)$ is the frequency of crossing the transition state due to thermal motion (h is Plank's constant), and C^0 and n are the standard state concentration and the order of the reaction respectively. $\gamma(T)$ is a term which accounts for barrier re-crossings, tunneling and any deviations from equilibrium. The physical picture from this expression is that the zero-point energy lowers an otherwise fixed free energy barrier with tunneling occurring at the top of the barrier to enhance the rate [209]. This expression extends the semi-classical treatment embodied in equation 6.3 and has been applied to AADH [197, 371] using semi-empirical quantum-mechanics/molecular mechanics (QM/MM) methods to calculate the tunneling correction with tryptamine substrate. Using this method a KIE of 30 in reference [371] which agreed, in magnitude if not precisely, with the experimental 55. The distance travelled by the tunneling proton was also found to be $0.59\,\text{Å}$ (i.e., well below its de Broglie wavelength). The temperature dependence of the KIE was not explored however, and it has been suggested [199, 207, 208, 216] that full-tunneling models are needed to explain the KIE and its temperature dependence.

The rate expression in full-tunneling models of Kuznetsov and Ulstrup [204, 214, 387] can be written as follows:

$$k = C(T)\frac{[V]^2}{\hbar}\sqrt{\frac{\pi}{\lambda k_B T}}\,e^{-(\Delta G^\circ+\lambda)^2/(4\lambda k_B T)}\int_0^\infty P(m,r)\exp\left[\frac{-E(r)}{k_\text{B}T}\right]\mathrm{d}r \tag{6.5}$$

(The following description of this equation is taken from reference [208].) The term outside of the integral represents a temperature dependent factor which accounts for the enzyme achieving a 'tunneling ready state' and is broken down as follows: $C(T)$ is a term representing the fraction of enzyme conformations that are able to go on to react; V is the coupling between the reactant and product wave-functions; λ is an energy associated with reorganising the active site ready for reaction; ΔG° is the free-energy difference between the reactant and product states; $\hbar$ is Plank's constant divided by 2π; k_B is Boltzmann's constant, and T is the thermodynamic temperature. This expression is the non-adiabatic rate expression where the first excited state (which are mixtures of vibrational and electronic states) play an important role, which is measured by V. Calculations have shown inclusion of these states important for the AADH reaction [216]. The difference between the effects captured by $C(T)$ and λ is a matter of scale: $C(T)$ corresponds to the metastable dynamics which brings the substrate-enzyme complex into a reaction ready state, while λ corresponds to sub-nanosecond rearrangement of bonds and the charge distribution into a tunnelling ready state. In this tunneling ready state the reactant and product states become degenerate and tunneling can occur with a probability described by the term inside the integral. $P(m,r)$ is the probability of tunneling occurring for a particle with a mass m, travelling over a distance r - the donor-acceptor distance (DAD). In the case of AADH, this is the difference between the reactant C—H bond and product O—H bond lengths. $\exp(-[E(r)/(k_{\mathrm{B}}T)])$ is the Boltzmann weight of each value of r. The physical picture is that thermal fluctuations of the enzyme align the reacting atoms into positions for which the reactant proton wave-function can overlap with the product wave-function and tunneling can take place [208, 387]. Fluctuations along the reaction coordinate sample values of the DAD which tunnel with a probability $P(m,r)$. As deuterium or tritium atoms have larger mass and shorter de Broglie wavelengths than hydrogen, $P(m,r)$ is peaked at smaller values of r, meaning the enzyme has to achieve smaller values of the DAD in order for the reaction to proceed [208, 387].

The advantage of the full-tunneling model (equation 6.5) over the tunneling correction (equation 6.4) picture is that it can account for both temperature dependent and independent KIEs [199, 208]. For temperature independent KIEs the enzyme/substrate complex achieves a tunneling ready state such that the DAD is small enough for tunneling to occur *without* the need for thermal fluctuations along r (or that those fluctuations do not have an appreciable energetic barrier, $E(r)$). For temperature dependent KIEs, the tunneling ready state does not have significant reactant/product wave-function overlap and significant fluctuations along the DAD are required, introducing a temperature dependence in the KIE.

Full-tunneling models have been used to explain the rates and KIEs in a variety enzymatic reactions. In their investigation of AADH, the authors of [216, 217] fit kinetic data with a tryptamine substrate to equation 6.5, along with analysis of short MD data, to identify a sub-picosecond long vibration which causes fluctuations in the DAD, allowing tunneling to occur. This vibration was identified as the rotation of the methylene group containing the donating carbon and transferring proton (C1, HI-2, HI-3 in figure 6.7). The energy barrier to this fluctuation, which in principle

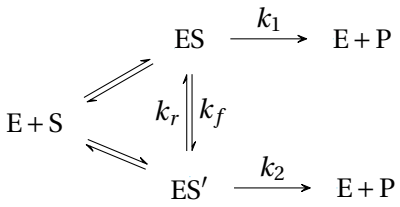


FIGURE 6.3: TWO CONFORMER TRANSITION STATE THEORY MODEL [221]. Two enzyme substrate conformers, ES and ES′, rapidly interconvert with rates k_f and k_r. Each conformer can react via distinct chemical pathways with rates given by k_1 and k_2.

should give rise to a temperature dependent KIE, was so low as to make the temperature effect indistinguishable from experimental results. Full-tunneling models have been used to rationalise the KIEs of the flavoenzymes morphinione reductase (MR), pentaerythritol tetranitrate (PETN) reductase (both reducing flavin the presence of nicotinamide coenzyme), as well as methylamine dehydrogenase (MADH) [218]. PETN, reductase like AADH, displays strongly temperature dependent reaction rates but negligible KIE temperature dependence, which the authors of suggest is due to the rigid and highly geometrically optimised active site of PETN reductase. This active site attains an optimal value of the DAD for tunneling (i.e., $P(m,r)$ is peaked at the configuration of the active site). MR by contrast (which is a close PETN reductase homologue), shows a KIE which is strongly temperature dependent. The authors rationalise this in terms of MR's more flexible structure which facilitates fluctuations in the DAD. This in turn allows the reaction in the less rigidly optimised active site of MR to proceed via tunneling. MADH in its reaction with methylamine shows similar behaviour to AADH, however, this was found to be substrate dependent, with ethylamine showing temperature dependent KIE. The authors suggest this was due the fact that MADH is highly specific for methylamine, hence the complex with ethylamine would be not optimised for tunneling to occur, requiring a thermal fluctuations to sample appropriate values of DAD. A similar picture arises in the study of mutants of isoleucine-14 (I14) in the active site dihydrofolate reductate (DHFR) [219]. The temperature dependence of the KIE for the wild type and three active site mutations were measured and were found to correlate inversely with the size of the replacement residue, i.e., the smaller the residue the larger the temperature depedence of the KIE. The smaller residues allowed a broader distribution of DADs to be sampled, which in turn required thermal activation to achieve tunneling of hydride in the rate determining step. Full-tunneling models have been applied in many other systems, see the reviews of reference [207, 208, 388] for more applications.

In contrast to the full-tunneling model, the authors of references [220, 221] showed that the temperature independence of the KIEs of AADH, MADH, DHFR and SLO enzymes could be explained using tunneling corrected TST by considering two reactive enzyme/substrate conformations as shown in figure 6.3. The two conformers ES and ES′ are in equilibrium and both are able to react. The rates of reaction from both conformers are k_1 and k_2, which are modelled using TST, equation 6.4. Assuming fast (compared k_1 and k_2) equilibration between the two conformer, an expression

for the KIE in terms of the various parameters of equation 6.4 was derived (i.e., $\Delta G^{TS,0}$ etc.). After fitting this expression to experimental data, they showed the fitted parameters were consistent with reasonable values for the parameters in k_1 and k_2, however, this has been disputed in reference [208]. The kinetic scheme in figure 6.3 has some similarity with the mechanism of AADH. As already mentioned, the rate limiting proton transfer can occur to either one of the aspartate oxyanions (OD1 or OD2, see figure 6.2) which provides a rational for identifying conformations which promote transfer to either one. However, it should be noted that in the most recent modelling [372] the reaction pathways to OD1 and OD2 both start with very similar conformations. In addition, previous simulations of AADH [197, 371] have not found any evidence for conformations favouring one pathway or the other.

6.1.4 Catalysis and protein dynamics

Conformational changes accompany the catalytic cycle of an enzyme as has been discussed in the introduction to this thesis and elsewhere [389, chapter 10][3, 225]; fast protein motions, such as the fast bond rotation in AADH discussed above (and other examples [208, 388]) are implicated in facilitating enzymatic reactions. The question of whether protein dynamics, broadly defined (i.e., not just long timescale conformational changes), provide the source of the enzyme's catalytic ability is a long-standing controversy [225, 387, 390–392]. To quantify enzyme catalysis a suitable reference reaction, without the enzyme present, must be defined and its rate, k_{ref}, compared to the rate of the enzymatic reaction k_{cat}. The role of enzyme dynamics to catalysis can then be quantified as its contribution to the ratio $k_{\mathrm{ref}}/k_{\mathrm{cat}}$. No measurement or calculation of k_{ref} for AADH has been performed and so a strict delineation of enzyme effects on catalysis cannot be made in this case. However, numerous definitions of 'dynamics' have been suggested and their potential role in catalysis explored [392] in other systems which can be a useful comparison for AADH. While the definitions differ slightly in their particulars, they all describe situations in which important motions are not well described by Boltzmann probabilities, i.e., non-equilibrium dynamics.

The principle definition of dynamical effects is through the transmission coefficient, the part of the pre-factor $\gamma(T)$ in equation 6.4 which accounts for non-equilibrium dynamics [390, 392]. Transmission coefficients less than 1 reflect motions of the protein not due to thermal fluctuations of the kind considered so far, coupling to the reaction coordinate and causing the system to re-cross the transition state back to the product state. However,no difference in the transmission coefficients between the solvated and enzymatic reactions have been found in, for example, haloalkane dehalogenase[228], thymidylate synthase [393] or catechol O-methyltransferase [394]. Another definition of dynamical effect is the 'intertial memory' proposal [392] which has been applied to adenylate kinase [226–228]. The authors of [390, 392] define this proposal as follows: excess kinetic energy upon transition from a non-reactive into the reactant conformation, is not dissipated but is rather channeled into the reaction coordinate, pushing the system into the product state. However, a follow up computational study using a multiscale modelling approach [229] allowing simualtions on the millisecond timescale

(the timescale of the reaction) found no coupling of the conformational transitions to the reaction coordinate [224]. It should be noted that the full-tunneling models described earlier have been described as promoting dynamical effects [392], however, as stated by Klinman and Kohen in reference [208] "... Marcus-like models neither require nor assume nonequilibrium dynamics".

6.1.5 Aims and objectives

The above discussion has shown that the conformational dynamics plays an important role in the reaction rate models which attempt to explain tunneling and the kinetic isotope effects in AADH and other enzymes with large tunneling components. In addition, multiple conformers have been implicated (although the evidence is far from conclusive) in non-equilibrium rate theories. It is clear then that building a model of the conformational dynamics of AADH will be an important contribution to these debates. Chapters 4 and 5 have demonstrated a framework for choosing an optimal set of hyperparameters (protein features, number of microstates, number of coarse-grained metastable states) for Markov state models. The aims of this chapter are thus two fold:

1. to arrive at a description of the conformational dynamics of AADH in its reactive state using Markov state models and the methods demonstrated in chapters 4 and 5.
2. to critically assess those methods using the 'real-world' example of AADH.

This chapter is structured as follows: section 6.2 describes the creation of a molecular dynamics data set and compares it to the most accurate computational study of the reaction mechanism. Section 6.3 uses the optimisation techniques from chapter 4 to create an optimal discretization of the MD data set. This is then coarse-grained in section 6.4 using an hidden Markov model with the number of metastable states determined using the ICL, as described in chapter 5. A summary and discussion of the implications of the final set of models is given in section 6.5, and section 6.7 concludes with a critical assessment of the modelling approach and discusses limitations.

6.2 Molecular dynamics

6.2.1 Simulation protocol

A PDB file of AADH was prepared by Dr Kara Ranaghan as described in section 6 at the start of this chapter. The atom types in TTW were changed to be compatible with the CHARMM-36 [395] force-field although the TTW parameters remained the same. The CHARMM package, version 42a2 [396] was used to create a protein structure file (PSF).

A solvation shell tracking the surface of the protein was created using the package Solvate (version 1.0) [397] which was modified by the author of this thesis to take into account the CHARMM extended PSF format for large systems. The water shell was $12.0\,\text{Å}$ thick, the maximum boundary curvature radius of the solvent surface was $100\,\text{Å}$, and 10 Gaussian functions were used to determine the solvent surface. This structure was further solvated to create a cubic simulation

cell of size $130\,\text{Å}$ using the package Visual Molecular Dynamics (VMD) (version 1.9.3) [398] with a boundary parameter equal to $1.2\,\text{Å}$. The size of the box was chosen so that the minimum distance between the enzyme and the edge of the simulation cell was $14\,\text{Å}$. The system was neutralized with VMD using sodium chloride to attain a concentration of $0.15\,\text{mol}\,\text{dm}^{-3}$. The minimization, heating and equilibration steps were performed in CHARMM with the OpenMM [399] plug-in using the CHARMM-36m [395] force-field. The electronic non-bonded forces in all steps were treated with partial mesh Ewald summation [336] with a cut-off of $14\,\text{Å}$, all other parameters were set to their default values. The SHAKE [400] algorithm was used throughout to constrain the bonds to hydrogen atoms. The minimization proceeded by first restraining all heavy atoms using a root mean square deviation (RMSD) restraint with a mass weighted force constant of $5\,\text{kcal}\,\text{mol}^{-1}\,\text{Å}^{-2}$ and then minimizing with 100 steps using the steepest descent algorithm and 3000 steps of adopted basis Newton-Raphson (ABNR) minimization. This was repeated three times, first limiting the restraint to all the heavy protein atoms, then the heavy protein backbone atoms, and finally with no restraint. The final unrestrained minimization proceeded with 5000 instead of 3000 steps of ABNR minimization.

The system was heated from $10\,\text{K}$ to $310\,\text{K}$ in steps of $25\,\text{K}$, with a harmonic restraint on the heavy protein backbone atoms with mass weighted force constant of $5\,\text{kcal}\,\text{mol}^{-1}\,\text{Å}^{-2}$. At each heating step $10\,\text{ps}$ of Langevin dynamics [231, 232] in a constant volume, constant temperature ensemble were run with a time-step of $2\,\text{fs}$ and collision frequency of $\gamma = 5\,\text{ps}^{-1}$.

The system underwent equilibration in two stages: i) restrained equilibration and ii) unrestrained equilibration. In the restrained equilibration stage $11 \times 20\,\text{ps}$ iterations of Langevin dynamics were run in an constant pressure, constant temperature ensemble (P=1atm, T=310 K), with a collision frequency $\gamma = 1\,\text{ps}^{-1}$, a Monte Carlo barostat with volume moves performed every $50\,\text{fs}$, and a time-step of $2\,\text{fs}$. On the first iteration the backbone atoms had a harmonic restraint with a mass weighted force constant of $10\,\text{kcal}\,\text{mol}^{-1}\,\text{Å}^{-2}$. On each subsequent iteration the force constant was reduced by $1\,\text{kcal}\,\text{mol}^{-1}\,\text{Å}^{-2}$ until the last iteration had no harmonic restraint. The unrestrained equilibration consisted of $200\,\text{ps}$ of equilibration run under the same conditions as before but with no restraints.

The simulation system was transferred from CHARMM to AMBER (version 16) [401] to make use of the improved user interface and post-simulation analysis tools. A single $100\,\text{ns}$ trajectory was produced in a constant volume, constant temperature ($T = 310\,\text{K}$) ensemble, using Langevin dynamics with a collision frequency of $\gamma = 5\,\text{ps}^{-1}$, and a time step of $2\,\text{fs}$. The non-bonded cut-off distance was reduced to $12\,\text{Å}$ and SHAKE [400] was used to constrain the hydrogen atoms. Coordinates and velocities were written to disk every $100\,\text{ps}$.

The coordinates and velocities at every $1\,\text{ns}$, $2\,\text{ns}$ ... $100\,\text{ns}$ were used to seed $100 \times 100\,\text{ns}$ new trajectories, run under the same conditions. These 100 trajectories constituted the AADH "production" data set.

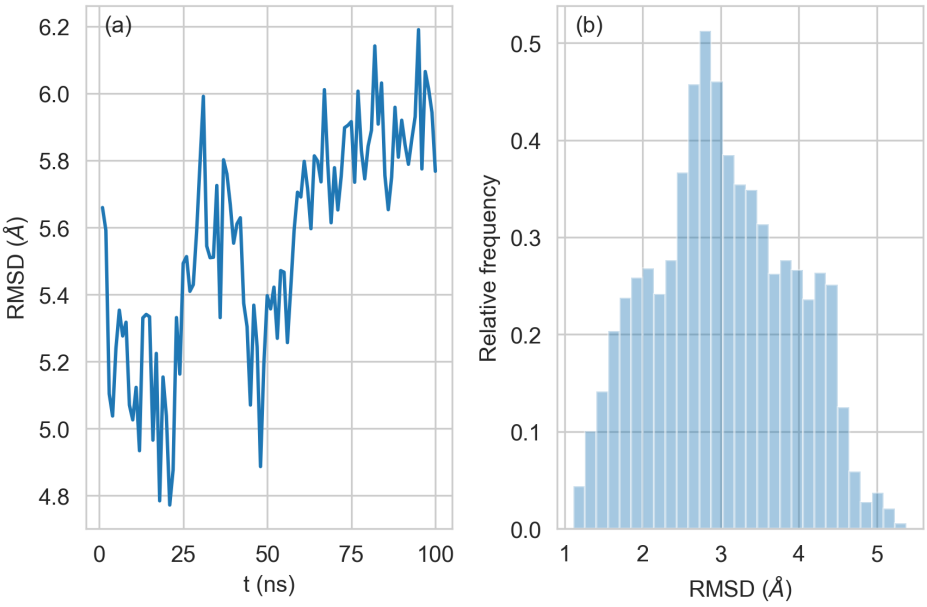


FIGURE 6.4: STRUCTURAL SIMILARITY OF SEED TRAJECTORY. A 100 ns trajectory, with coordinates saved every 1 ns was used to seed 100 production trajectories. Panel (a) shows the α-carbon RMSD of the seed trajectory relative to the crystal structure, (b) shows the distribution of all pairwise α-Carbon RMSD within the seed trajectory.

6.2.2 Modelling disulphide bridges

A missing disulphide bridge was found in the preparation of the initial structure *after* completion of the simulations. The disulphide bridge between Cys81 and Cys113 on the H chain (but not on the D chain) had not been created, instead the two thiol groups were left unoxidized. This introduced differences in the structures of the active sites which are described in full in section 6.2.6. The conclusions of this analysis are given in section 6.2.8.

6.2.3 Correlation between seeding conformations

In order to check the degree of correlation between the structures used to seed the 100 production simulations, the structural similarity between the seed frames was calculated and is shown in figure 6.4. Panel (a) shows the α-carbon RMSD, relative to the crystal structure, which shows a persistent, large deviation from the crystal structure and serial correlation between initial frames for $t > 60\,\text{ns}$. Panel (b) shows the distribution of α-Carbon RMSD values between each pair of initial frames (i.e., there are $\frac{1}{2} \cdot 100 \cdot 99 = 4950$ values in the histogram). This demonstrates a range of differences between the initial conformations: 50 % of the initial structures have an RMSD of 3.0 Å or larger. While there is a degree of correlation between the seed conformations, this was deemed small enough to keep all sampled trajectories in the remaining analysis.

6.2.4 Structural stability

Each trajectory was checked for structural stability over the course of the simulations by calculating the RMSD of the α-carbons atoms relative to the crystal structure, this is shown in figure 11.1. The majority of the trajectories showed a stable RMSD as a function of time, with $95\,\%$ of the frames remaining within $4.5\,\text{Å}$ to $6.4\,\text{Å}$ of the crystal structure. However, seven of the trajectories (24, 27, 30, 42, 78, 87 and 97) had an RMSD trajectory which increased over time, leaving the upper $95\,\%$ bound by at least the final frame, shown in figure 11.1. The source of the increase in RMSD was not due to changing secondary structure (figure 11.2) but rather due to the high fluctuations in the tail residues, see figure 11.3. The production trajectories thus demonstrated adequate structural stability.

6.2.5 Enzyme flexibility

In order to understand the origin of the large values of RMSD seen in figure 6.4, a breakdown of the RMSD on the seed trajectory at $95\,\text{ns}$ was performed (this frame was chosen as it had the highest RMSD of $6.2\,\text{Å}$). The deviation of α-carbon positions of each residue for frame $95\,\text{ns}$ relative to the crystal structure was calculated (the values are plotted in figure 11.4). The contributions to this large value of RMSD arise primarily from the residues at the tails of chains D and A, as well as from loop residues 92–108 in chain D. Removing these tail residues reduced the RMSD to $5.2\,\text{Å}$.

The loop residues 92–108 in chain D (and to a lesser extent chain H) show a large deviation due to a significant conformational change, shown in figure 6.5. This figure shows the backbone of the D chain only, with loop 92–108 coloured green, the two active site residues (Asp128 and TTW109) shown in orange, and the remaining protein in blue. On the left hand side is the crystal structure, on the right hand side is the snapshot at $95\,\text{ns}$, while view 1 and view 2 correspond to two different camera angles. The diameter of the backbone in all images corresponds to the size of the deviation from the crystal structure of the snapshot at $95\,\text{ns}$. As can be seen, the loop moves across the protein, away from the Asp128 residue. Removing these loop residues the calculation, the RMSD is reduced further to $4.9\,\text{Å}$, indicating that the rest of the protein in moderately aligned to the crystal structure.

To understand the flexibility of the enzyme over the whole of the simulation data set (rather than the snapshot presented in figure 6.5), the root mean square fluctuation (RMSF) per residue was calculated and shown in figure 6.6. The picture presented here does not differ significantly from the that implied in the preceding analysis. There residues at the tails of each chain, particularly in chain A, exhibit extreme flexibility. The loop 92–108, in both chain H and chain D shows the next highest degree of conformational flexibility. Residues 110–126 correspond to a β-hairpin structure which also shows a higher than average degree of flexibility, especially in chain D. However, the tip of this hairpin, which is the most flexible, is greater than $12\,\text{Å}$ from the active site in the crystal structure.

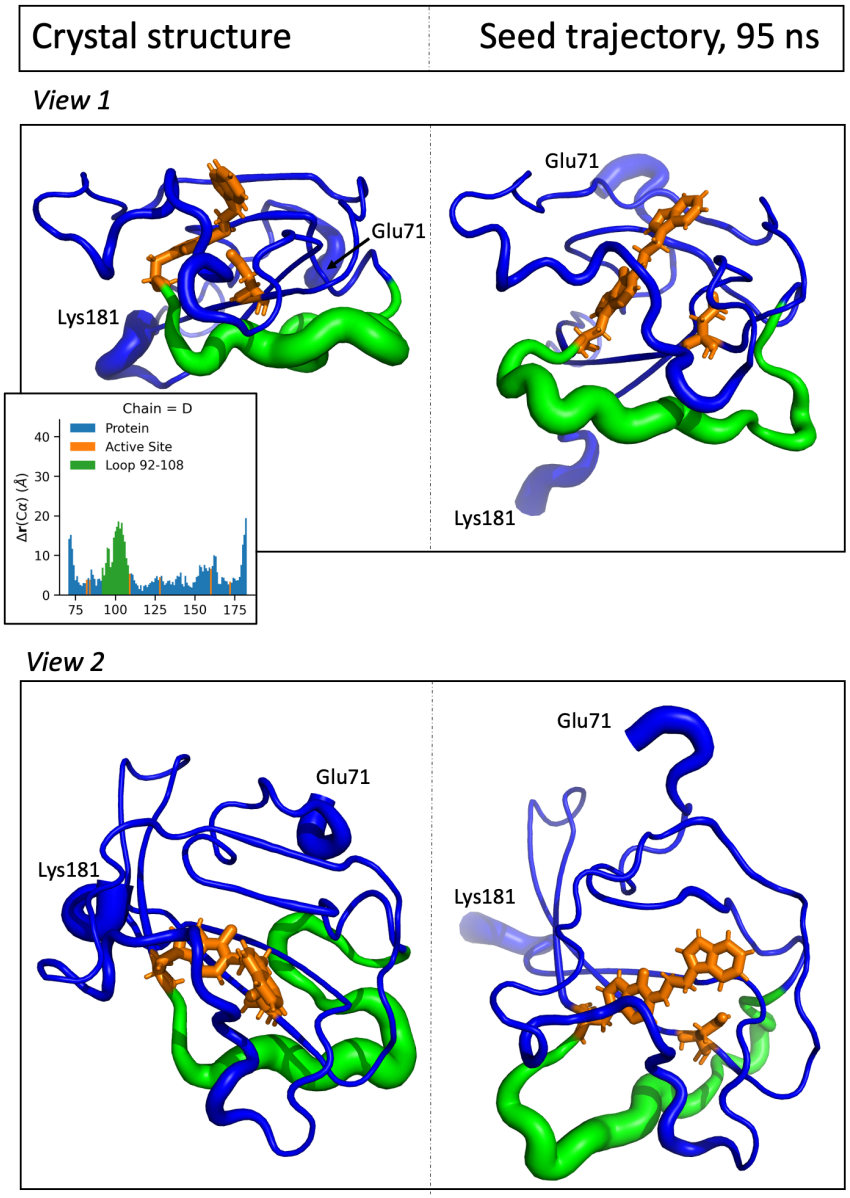


FIGURE 6.5: CONFORMATIONAL CHANGE IN LOOP RESIDUES 92–108 OF CHAIN D. The crystal structure (left) and the seed trajectory at 95 ns (right) are shown from two different camera orientations (view 1 and view 2). The diameter of the backbone is proportional to the deviation of the α-carbon from the crystal structure. The chart inset shows the deviation by residue of chain D, taken from figure 11.4. The loop residues 92 to 108 are highlighted in green, two active site residues TTW109 and Asp128 are shown in orange (the colour coding is consistent between the inset and the depicted conformations, except that four of the active site residues are not shown). The N-terminus (Glu71) and C-terminus (Lys181) are labelled. The 22 tail residues, unresolved by the crystal structure, are missing from the N-terminus.

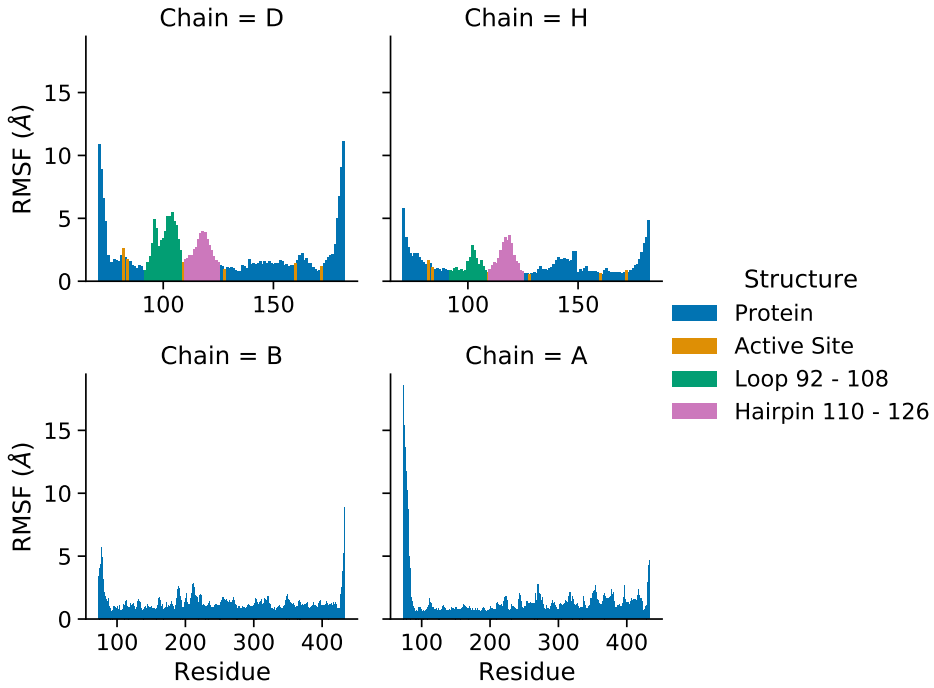


FIGURE 6.6: ROOT MEAN SQUARE FLUCTUATION PER RESIDUE OF AADH. Each panel shows the root mean square fluctuation of (clockwise from top left) chains D, H, B and A. Residues in loop 92–108 are shown in green, residues in the β-hairpin 110–126 are shown pink, the active site residues are in orange.

6.2.6 Definition and comparison of active sites

The active site of the enzyme was defined in the same way as references [197, 371, 372] in chains D and H: Ala82, Asp84, TTW109, Asp128, Trp160 and Thr172. The TTW residue is the TTQ prosthetic group after reaction with the tryptamine substrate to form the Schiff base intermediate. Hereafter any reference to TTQ will refer to the portion of TTW coming from TTQ originally and not the unreacted prosthetic group. The crystal structure of the active site is shown in figure 6.7.

The structures of the two active sites were also compared to the crystal structure. The distribution of the heavy atom RMSD, relative to the crystal structure, is shown in figure 6.8. This shows that the H active site is more structurally similar to the crystal structure than the D active site. This could be associated with the difference in the conformations of the loop residues 92–108 between chain D and H in figure 11.4 and with the missing disulphide bond in chain H. However, this link has not been investigated.

To understand this difference between the active sites further and to explore the similarities between these simulations and previous work [372], the distribution of important interatomic distances were calculated and compared. Figure 6.9 shows these bond distributions in blue for the D active site and in orange for the H active site. The black vertical lines with labels are the QM/MM

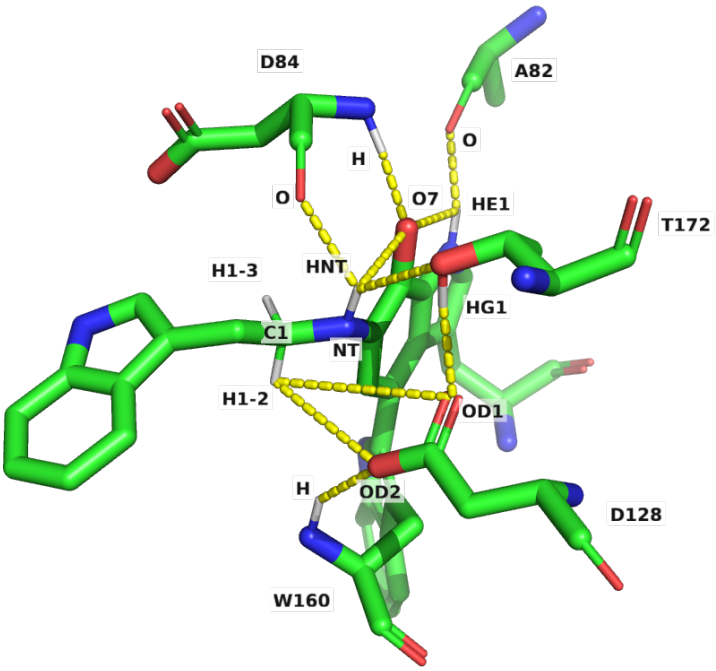


FIGURE 6.7: CRYSTAL STRUCTURE OF THE ACTIVE SITE OF AADH. This definition is taken from reference [372] and consists of the following residues: Ala82, Asp84, TTW109, Asp128, Trp160 and Thr172. The yellow dashed lines show the important stabilizing hydrogen bonds as well as the H—O distances involved in the rate limiting step. The reactive hydrogen atoms are labelled HI-2 and HI-3. The acceptor oxygen atoms are labelled OD2 and OD1 which correspond to O1 and O2 in reference [197] respectively. All other hydrogen atoms are hidden. Backbone carbonyl bonds may appear as single bonds due to the camera angle.

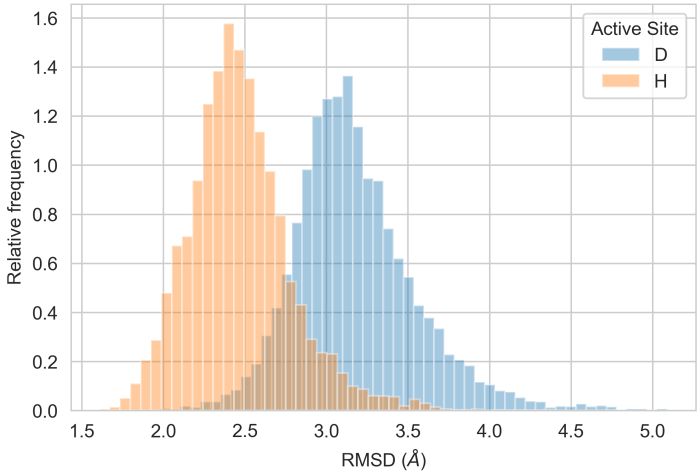


FIGURE 6.8: COMPARISON OF THE ACTIVE SITE IN CHAINS D AND H. The distribution of heavy atom RMSD, relative to the crystal structure of the active sites in chain D (blue) and H (orange).

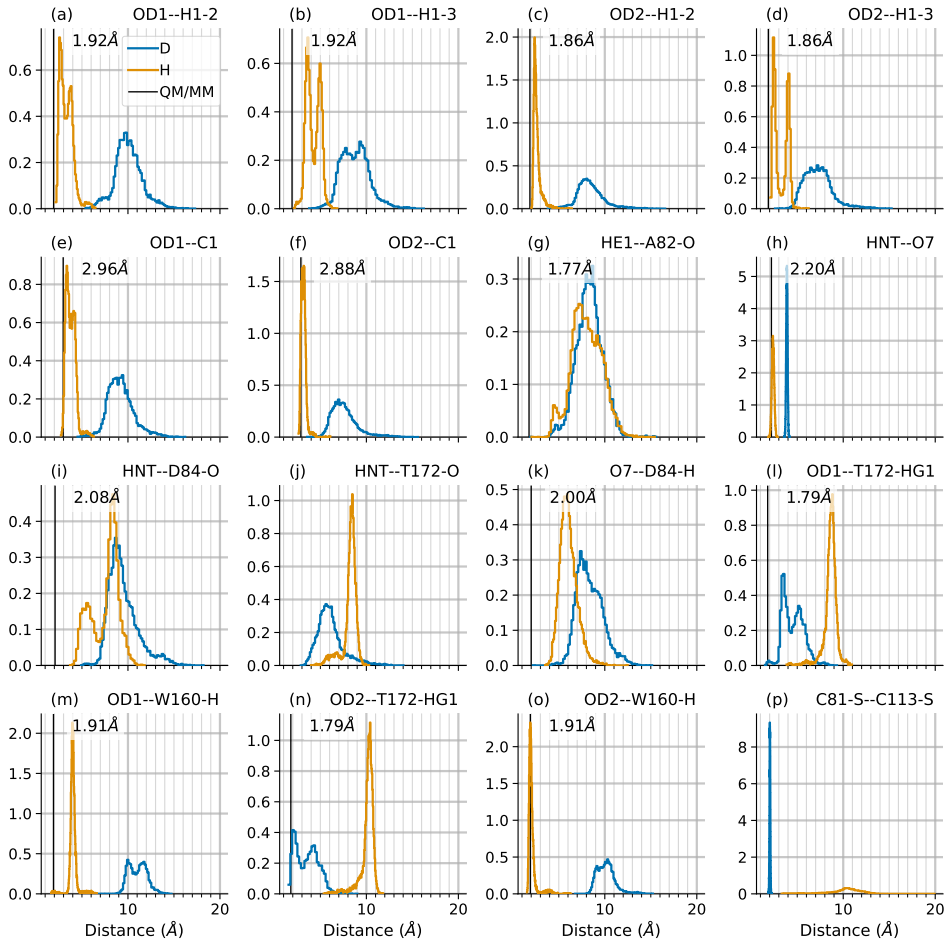


FIGURE 6.9: DISTRIBUTION OF IMPORTANT BOND DISTANCES IN THE ACTIVE SITE. Panels (a) - (d) show the four combinations of acceptor ion (O1, O2) - proton (H1-2, H1-3) distances; panels (e) and (f) are the two donor (C1) acceptor (O1, O2) distances; the remaining panels are the hydrogen bonds in the active site.

interatomic distances in the reactant state[2], taken from table 3 of reference [372]. Where possible the atom labels have been kept the same as those in reference [372], the exceptions are the atoms directly involved in bond breaking and formation. The correspondence between the interatomic distances in reference [372] and figure 6.9, and their description are as follows:

1. OD1/OD2—H1: The bond being formed. These correspond to all four combinations of distances between OD2/OD1 (respectively) and H1-2, H1-3. Shown in panels (a) through (d).
2. OD1/OD2—C1: The donor/acceptor distance. These are the two combinations of distances between OD2/OD1 (respectively) and C1. Shown in panels (e) and (f).
3. HE1—A82-O: inter-residue hydrogen bond between HE1 of the TTQ prosthetic group and the Ala82 backbone oxygen atom, shown in panel (g).

[2]An average over the pathways to OD1 and OD2 are taken where appropriate.

4. HNT—O7: intra-residue hydrogen bond in the TTW residue. Shown in panel (h).
5. HNT—D84-O: inter-residue hydrogen bond between the TTW residue and the backbone oxygen atom of the Asp84 residue. Shown in panel (i)
6. HNT—T172-O: inter-residue hydrogen bond between TTW and the backbone oxygen atom in Thr172. This is not described in reference [372] but is included because of its potential to form a hydrogen bond in certain conformations. Shown in panel (j).
7. O7—D84-H: inter-residue hydrogen bond between the TTW residue and the backbone amide hydrogen atom of the Asp84 residue. Shown in panel (k).
8. OD1—T172-HG1/W160-H: The inter-residue hydrogen bond between the acceptor oxygen atom OD1 of D128 and hydrogen atoms on the Thr172 and Trp160 residues respectively. Shown in panels (l) and (m).
9. OD2—T172-HG1/W160-H: The inter-residue hydrogen bond between the acceptor oxygen atom OD2 of D128 and hydrogen atoms on the Thr172 and Trp160 residues respectively. Shown in panels (n) and (o).
10. C81-S—C113-S: The sulphur-sulphur distance corresponding to the missing disulphide bond in the H chain. Shown in panel (p).

In general, the bond lengths in the D active site show greater variance and do not overlap with those in the H active site. The main exception to this are the from the Cysteine residues 81 and 113. These have clearly drifted apart due to the lack of disulphide bond: the distribution of the S—S distances in the H active site, shown in panel (p) varies between $3\,\text{Å}$ to $20\,\text{Å}$ compared to the effectively fixed distance of $\simeq 2\text{Å}$ in the D active site.

The O—H distances are smaller, closer to the QM/MM values, and show less variation in the H active site compared to the D active site by a significant margin (panels (a) - (d)). The distances in the H site are almost all less than $5\,\text{Å}$ (within $\simeq 3\text{Å}$ of the QM/MM values), where almost all the D active site values are between $5\,\text{Å}$ to $10\,\text{Å}$. The donor/acceptor distances, C—O (panels (e) and (f)) show a similar story. The closest hydrogen atom to both acceptor oxygen anions in the H active site is H1-2, the differences in the D active site are less obvious. These differences are due to the crystal structure preparation, missing disulphide bond, and potentially due to conformational changes in loop residues 92–108, as: i) there is no difference between the active sites in the crystal structure (the difference in heavy atom RMSD $< 0.01\text{Å}$); and ii) the broken disulphide bond is incompatible with the pathways determined in references [371, 372].

There is no conclusive similarity between the QM/MM results and these simulations with respect to the stabilization of OD1 and OD2 by Thr172 and Trp160 by hydrogen bonds. This is important as it is these hydrogen bonds which help define the difference in between the hydrogen abstraction pathways. The orientation of Trp160-H, OD2, OD1 and Thr172-HG1 in the crystal structure and QM/MM is approximately linear. For OD1 to be hydrogen bonded with Thr172, the bond lengths would be ordered OD1—Thr172-HG1 < OD1—Trp160-H, i.e. panel (l) < panel (m). This is true for D but not for H. While for OD2 to be hydrogen bonded with Trp160, the bond lengths would be

ordered OD2—Thr172-HG1 > OD2—Trp160-H, i.e. panel (n) > panel (o). This is true for H but not for D. In the D active site, Asp128 has moved away from Trp160 entirely but remained close to Thr172 as seen by comparing the distributions in panels (m) and (o) with panels (l) and (n). Noting that TTW is covalently bonded to Trp160, this is then consistent with the picture in panels (a) through (f) where OD1 and OD2 are between $5\,\text{Å}$ to $15\,\text{Å}$ from the relevant atoms on TTW.

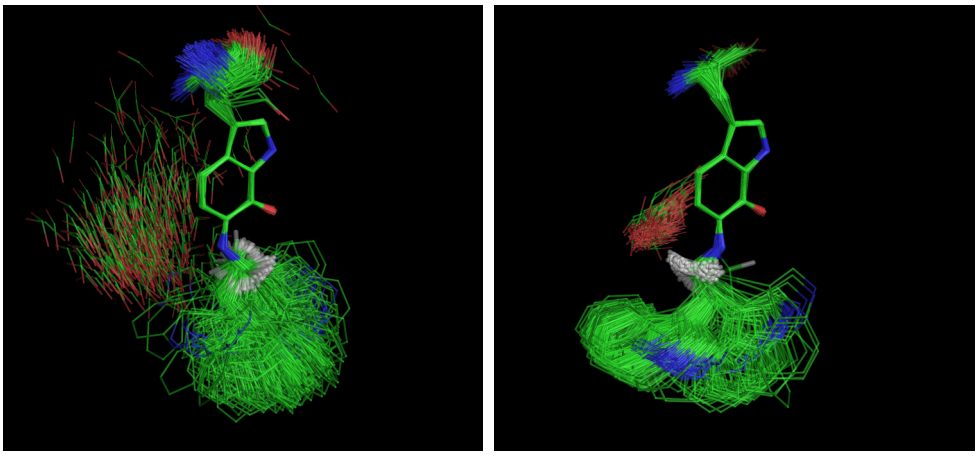
The TTW intra-residue hydrogen bond, HNT—O7, shows good agreement with the QM/MM value in the H active site, while in the D active site it is larger by almost $2\,\text{Å}$. The two distinct HNT—O7 bond lengths define whether the NT—C1 bond is either syn or anti the C—O7 carbonyl bond in the tryptophan ring system of TTQ. The anti conformation, with has the shorter HNT—O7 distance, is adopted in active site H and in the crystal structure as shown in figure 6.7. Here the NT—HNT bond points in the same direction as the C—O7 bond and forcing the NT—C1 bond into the anti-conformation. The syn conformation has the has the NT—HNT bond pointing in the opposite direction, forcing the NT—C1 bond to eclipse the the C—O7 carbonyl bond.

The syn and anti conformations allow radically different conformational states to be accessed as demonstrated in figure 6.10. Each panel shows 500 configurations, selected evenly across all trajectories, for the D (panel (a)) and H (panel (b)) active sites. The entire TTW residue and carbonyl group of Asp128 is shown and the conformations have been aligned to the tryptophan part of the TTW residue. The only two hydrogen atoms shown are H1-2 & H1-3 and are coloured white. The syn conformation of the D active site clearly demonstrates a 'looser' set of conformations with the C1-H1 bond pointing away from the acceptor Asp128 residue, although whether this is due to missing disulphide bond or loop 92–108 has not been determined. The anti conformation of the H active site shows a 'tighter' set of conformations with the C1-H1 bond pointing towards the Asp128 residue.

In summary, the two active sites show distinct differences in their accessible conformations. The two sites are differentiated in three main ways. First, the relevant interatomic distances and hydrogen bonds between TTW and Asp128 are more consistent with the QM/MM results in the H than in the D active site. Second, the orientation of the NT—C1 bond is syn the C—O7 carbonyl bond in the D active site, and anti in the H active site and QM/MM results. Third, the H site is more constrained with less variation in the available conformations compared to the D active site. The apparent greater compatibility of the H active site with QM/MM results is surprising given the missing disulphide bridge occurs in the H active site.

6.2.7 Donor-acceptor distance

A modified definition of the donor-acceptor distance (DAD) will be utilised throughout this chapter to discuss the degree of 'reactivity' of different conformations. The modified DAD is defined as the distance between C1 of the tryptamine in TTW109, which donates the proton, and either OD1 or OD2 of Asp128 which accepts the proton during the course of reaction, C1—OD1/2 (these distances are in panels (e) and (f) of figure 6.9). This is related but different to the definition in



(a) Active site D (b) Active site H

FIGURE 6.10: CONFORMATIONS OF THE TTW RESIDUE IN THE D AND H ACTIVE SITES. Five conformations were taken from each trajectory at intervals of 20 ns and aligned along the heavy atoms of the trypotophan part of the TTW residue. The hydrogen atom shown is the donor atom, all other hydrogen atoms are hidden.

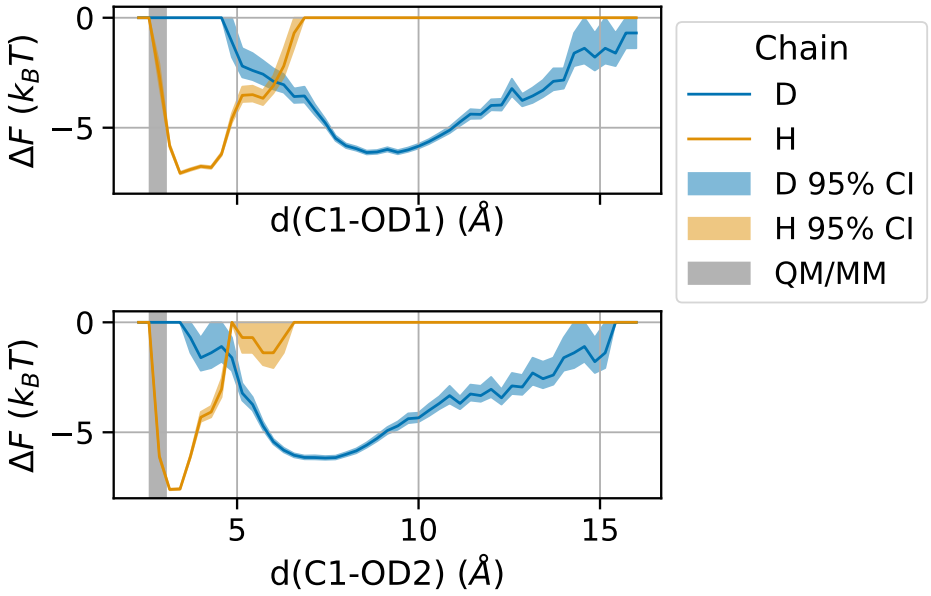


FIGURE 6.11: APPROXIMATE FREE ENERGY OVER THE DONOR-CARBON–ACCEPTOR-OXYGEN DISTANCE. The free energy was estimated by from a histogram (h_i) of bond distances with $i = 1,\ldots,50$ equally spaced bins over the range 2 Å to 16 Å with the formula $\Delta F_i = -\log(h_i)$. 95 % confidence intervals were estimated from 1000 bootstrap samples. Panel (a) shows the distance to acceptor OD1 and panel (b) shows the distance to OD2. Blue and green lines show the distances in chain D and chain H respectively. The grey shaded area shows the range of these distances as simulated in reference [372]. Black crosses and labels mark the distance corresponding to the free energy minima.

previous work [216] which is the distance travelled by the proton over the course of the reaction. As the reaction is not modelled in these simulations, this definition is not appropriate.

In the QM/MM modelling of reference [372] the value of the C1—OD1/2 distances in the reactant state are $2.96\,\text{Å}$ and $2.88\,\text{Å}$; in the transition state are $2.64\,\text{Å}$ and $2.59\,\text{Å}$; and in the product state are $3.04\,\text{Å}$ and $2.94\,\text{Å}$. Thus, over the course of the reaction pathway estimated in reference [372] these distances remain approximately constant (similar results are found in reference [197], figure 3).

In the simulations presented in this chapter, conformations will be considered reactive if they achieve a DAD of $\sim 3\,\text{Å}$. This ignores the both the orientation of the hydrogen atom and the role of other stabilising interactions in creating a potential energy surface conducive to tunneling. It may be, therefore, that defining reactivity just in terms of these values of the DAD will falsely identify conformations as reactive when, in reality, they are not. However, focusing on this definition of reactivity in terms of the DAD alone, is justified for the following reasons. First, including the hydrogen atom position in the definition of reactivity in a molecular mechanics force-field simulation is problematic because the bond distance is frozen in the SHAKE algorithm employed here and so not modelled accurately. Second, an appropriate value of C—OD1/OD1 distance is a necessary (albeit not sufficient) condition for achieving a reactive state. Third, to validate any reactive conformations found in this work, further QM/MM modelling will be necessary which will likely change many of the interactions and forces on atoms in active site (in particular the orientation of hydrogen atoms). Identification of reaction conformations can only be an approximate classification and thus focusing on a liberal definition of reactivity is justified. Fourth, the goal of the analysis here is to explore the possibility of other conformations which may be reactive. While the many interactions observed in reference [372] may be sufficient for the reaction to occur, they may not be necessary.

The approximate free energy pathways over the DAD are shown in figure **??** to both OD1 (panel (a)) and OD2 (panel (b)) for chain D (blue) and chain H (orange). For comparison, the ranges of these distances over the reaction pathway from the QM/MM modelling of reference [372] are shown as shaded areas. The distances sampled in chain H show a free energy minima just outside the reactive range ($3.3\,\text{Å}$ and $3.0\,\text{Å}$ to OD1 and OD2 respectively). Chain D show free energy minima more than double the reactive distance. This is in line with the analysis of the previous section which showed the conformations in the H chain are more inline with the crystal structure and the QM/MM modelling of reference [372].

6.2.8 Summary of molecular dynamics data

$10\,\mu\text{s}$ of molecular dynamics trajectories, seeded from 100 different conformations, has been simulated. Some of the seed conformations show similarity in their structures as measured by their RMSD from the crystal structure (figure 6.4 panel (a)) and their RMSD from one-another (figure 6.4 panel (b)). There are two regions of high flexibility (aside from the tail residues) next to the active site, loop 92–108 and a β-hairpin 110–126 (figure 6.6). The loop undergoes a significant conformational

change, an example of which is shown in figure 6.5, while the β-hairpin also shows flexibility at its tip, albeit at a distance of approximately $12\,\text{Å}$ from the active site. The active site residues do not show higher than average flexibility.

There are significant differences between the two active sites both in terms of their RMSD relative to the crystal structure (figure 6.8) and along key interactions (figure 6.9). This is due to the missing disulphide bridge between residues Cys81 and Cys113 in chain H. The active site in chain H surprisingly shows more similarity to the crystal structure but the missing disulphide bridge invalidates this data. Given this, the models of conformational dynamics estimated in this chapter will only use data from the active site in the D chain.

6.3 MSM optimisation

The aim of this section is to create an optimised Markov state model of the active site of AADH using Bayesian optimisation and response surface methods developed in chapter 4. In addition, to test the robustness of the modelling procedure a sensitivity analysis based on the response surface of the MSM will be produced. This section proceeds in four steps: first in subsection 6.3.1, a suitable value of the Markov lag time τ_{M}, and number of slow relaxation processes, r will be estimated using a reference MSM. These will be used to specify the MSM and the VAMP-2 score so that in the second step, subsection 6.3.2, the MSM response surface can be modelled as a Gaussian process (GP). Third, using this model of the response surface Bayesian optimisation will be used to determine the optimal set of hyperparameters, these will constitute the base case model. In the fourth step, subsection 6.3.4, a number of alternate models will be proposed ('sensitivity 1, 2' etc.) using knowledge of MSM response surface and eigenvalue spectrum. All calculations were performed using the packages cited in chapter 4.

6.3.1 Estimating the Markov lag-time and number of metastable states

Suitable values of the Markov lag-time, τ_{M}, and the number of dominant relaxation processes, r, were estimated by inspection of the eigenvalue spectrum of a reference MSM. The following results are for the D active site only as the missing disulphide bond in the H active site rendered these trajectories invalid.

The trajectories were discretized by first projecting the cartesian coordinates onto a set of features, applying TICA to reduce the dimension of the feature space and then clustering the frames into a small set of discrete states. The features used were the bond lengths identified in reference [372] i.e., those whose histograms are depicted in panels (a) - (g), (i), (k), (l), (o) of figure 6.9. The intramolecular hydrogen bond in panel (h), originally included in reference [372], was excluded due to its small variance. The trajectories were strided to give a trajectory time-step $0.1\,\text{ns}$. TICA was applied with a lag time of $\tau = 1\,\text{ns}$ and the number of components set so as to retain $95\,\%$ of the kinetic variance. This equated to retaining $m = 10$ components. k-means clustering was used to

cluster the trajectories into $n = 316$ microstates states. The number of microstates was chosen as the square root of the number of observations, inline with the heuristic in reference [246].

The Markov lag time must be chosen large enough so that the Markov assumption holds. This equates to the implied timescales:

$$t_i = -\frac{\tau_{\mathrm{M}}}{\ln|\lambda_i|},$$

being independent of τ_{M}. However, the eigenvalue spectrum will be heavily influenced by choice of hyperparameters, in particular the protein feature. So a suitable lag time for this particular set of hyperparameters may not prove suitable for another set. In other words, a suitable lag time cannot be chosen independent of the hyperparameters, but the hyperparameters cannot be optimised without specifying a lag time. The same reasoning also applies to the number of dominant eigenvalues used in the VAMP-2 score.

A way out of this circular reasoning problem can be found by noting the following. First, the purpose of specifying the lag time and number of dominant processes is to provide a starting point from which the hyperparameters of the MSM can be optimised. It is therefore only necessary that the choice of τ_{M} do not affect the optimisation, rather than provide a strictly valid MSM specification. The lag time and number of dominant relaxation processes has only a small effect on the value of the VAMP-2 score as demonstrated by figure 6.12. The first five implied timescales are shown in panel (a) for τ_{M} in the range $0.1\,\mathrm{ns}$ to $5\,\mathrm{ns}$ and panel (b) for τ_{M} in the range $0.1\,\mathrm{ns}$ to $50\,\mathrm{ns}$. Underneath, in panels (c) and (d) are shown the VAMP-2 scores with r ranging from 2 to 5, which correspond to successively including the implied timescales shown in panels (a) and (b). The timescale t_2 (shown in blue) appears to become independent of the lag time from approximately $12\,\mathrm{ns}$. however, as panels (c) and (d) show, the lag time has little effect on the VAMP-2 scores which vary by less than $\pm 4\,\%$ from their initial values over all lag times. The value of r also has little effect on VAMP-2 scores, at least up to $\tau_{\mathrm{M}} = 15\,\mathrm{ns}$ where their relative values start to diverge. Third, a value of τ_{M} and r used for inference can be determined after optimisation through appropriate sensitivity analysis. A value of $\tau_{\mathrm{M}} = 2\,\mathrm{ns}$ was chosen based on the relatively small variation of the VAMP-2 score (panel (c) of figure 6.12) and the larger number of observations that a small value affords.

There were no significant gaps in eigenvalues at $\tau_{\mathrm{M}} = 2\,\mathrm{ns}$ shown figure 6.13 panel (a). So the number of dominant processes was determined by inspection of the relative gaps in the timescale spectrum shown in panel (b). The largest timescale separation was between the fourth and fifth relaxation process, suggesting a value of $r = 4$.

An alternative analysis was performed using a TICA lag time of $\tau = 10\,\mathrm{ns}$ with $95\,\%$ of the variance ($m = 8$ components) retained with same number of cluster centres, $n = 316$. The implied timescale plots are shown in figure 11.6 and do not change the Markov lag time suggestion of $2\,\mathrm{ns}$. The ratio of eigenvalues and implied timescales are shown in figure 11.5. There was no clear separation in either the eigenvalues or the implied timescales and so the value of $r = 4$ from reference MSM was retained.

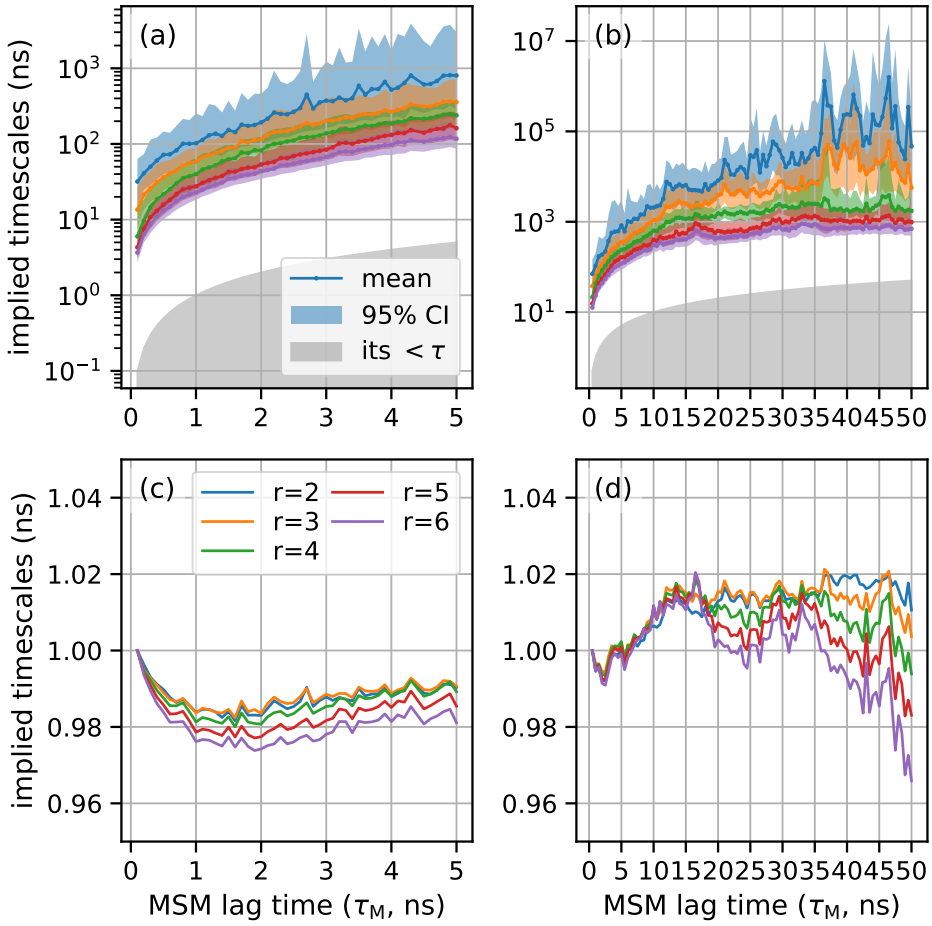


FIGURE 6.12: IMPLIED TIMESCALES AND VAMP-2 SCORES OF THE REFERENCE MSM. Panels (a) and (b) show the implied timescales, and panels (c) and (d) shows the relative VAMP-2 scores for the reference MSM with: $\tau_M = 2\,\text{ns}$, TICA lag time of $\tau = 1\,\text{ns}$, 95 % of the kinetic variance/$m = 10$ TICA components retained, and $n = 316$ microstates. Panel (a) shows the first five implied timescales for τ_M =0.1 ns–5 ns, panel (b) shows the first five implied timescales for τ_M =0.1 ns–50 ns. The solid lines and coloured shaded areas are the mean and 95 % credible intervals respectively, estimated using MCMC with 500 posterior samples. The grey shaded area is the region for which the implied timescales are smaller than the lag time. Panel (c) and (d) show the VAMP-2 scores, scored on the first 2 to 5 eigenvalues for the same ranges. The VAMP-2 scores are indexed to their value at $\tau_M = 0.1\,\text{ns}$. The colour coding is consistent between the implied timescale plots ((a) and (b)) and VAMP-2 plots ((c) and (d)). e.g. the blue line in (c) and (d) is the VAMP-2 score with two eigenvalues ($r = 2$) while in (a) and (b) blue is the second implied timescale, t_2.

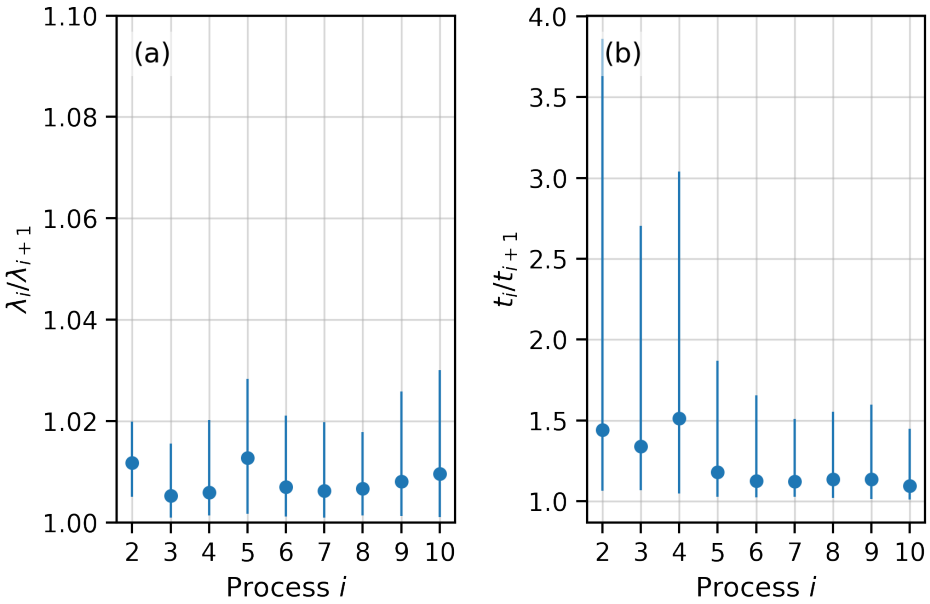


FIGURE 6.13: RATIO OF SUCCESSIVE EIGENVALUES AND IMPLIED TIMESCALES OF THE REFERENCE MSM. Panel (a) shows the ratio of eigenvalues and panel (b) the implied timescales for the reference MSM with: $\tau_{\mathrm{M}} = 2\,\mathrm{ns}$, TICA lag time of $\tau = 1\,\mathrm{ns}$, 95 % of the kinetic variance/$m = 10$ TICA components retained, and $n = 316$ microstates. Parameters were estimated using MCMC with 1000 posterior samples. The blue dots and error bars are the mean and 95 % credible intervals respectively.

6.3.2 Response surface

The response surface of an MSM for the active site of AADH was estimated using the same method as alanine dipeptide in chapter 4. A summary of this method applied to AADH is as follows. The MD data used were the 100 coordinate trajectories of the six residues of the D active site of AADH defined in figure 6.7. These trajectories were pre-processed into discrete states by first projecting onto a continuous feature χ; applying TICA with a lag time of τ and retaining m independent components $(\mathrm{IC}1,\ldots,\mathrm{IC}m)$; then clustering into n discrete microstates using k-means clustering. Collectively these are the MSM hyperparameters $\mathbf{x} = (\chi, \tau, m, n)$. The range of possible values of $\mathbf{x}$, the hyperparameter search space, are shown in table 6.1. The (ϕ, ψ, χ) feature was the usual backbone and residue dihedral angles but augmented with the six dihedral angles joining the two 9-membered rings in the TTW residue. 500 different sets of hyperparameters were sampled from the search space (100 per feature). To each value of $\mathbf{x}$ an MSM was fitted with a lag time of $\tau_{\mathrm{M}} = 2\,\mathrm{ns}$ and the response, y, measured with VAMP-2 with $r = 4$ eigenvalues using 20 iterations of 50:50 shuffle split (algorithm 1). The 500 pairs of hyperparameter/response measurements constituted the hyperparameter trial data set, $\mathscr{D}_{500} = \{(\mathbf{x}_i, y_i),\ i = 1,\ldots,500\}$. The response surface $f(\mathbf{x}; \mathscr{D}_{500})$, was modelled as a Gaussian process (GP), with covariance kernel and input warping selected using

Hyperparameter	Type	Range	Dim.	Details
Feature, χ	Categorical	(1) (ϕ, ψ, χ)	116	
		(2) $\lvert\mathbf{r}_1 - \mathbf{r}_2\rvert$	2346	Heavy atom interatomic distances
		(3) $C_\alpha - C_\alpha$	15	α-carbon contacts
		(4) $X - X$	15	Heavy atom contacts
		(5) RMSD	1	Heavy atoms only
TICA lag time, τ	Integer	1 ns, 1.1 ns ... 100 ns		
TICA components, m	Integer	1, 2 ... 20		
Cluster centres, n	Integer	10, 11 ... 1000		Clustered using k-means clustering

TABLE 6.1: HYPERPARAMETER SEARCH SPACE FOR AADH. TICA was applied to every feature except RMSD. Dihedral angles, θ, were given $(\sin(\theta), \cos(\theta))$ representations to account for their periodicity. The Cartesian coordinates were first aligned to a single, randomly chosen, trajectory frame so that feature (5) did not include spurious rotational or translational motion. The number of dimensions, 'Dim.', refers to the number of individual feature variables created by χ.

a combination of the mean standardized log-loss (MSLL) and the standardized mean square error (SMSE).

The hyperparameter trial data set, $\mathscr{D}_{500}$, is shown in figure 6.14. The mean test response is shown in blue and the difference between the training and test response shown in orange. Not all trials were successful (for example, if the number of TICA dimensions retained, m, exceeded the number of dimensions of the feature, χ) and these trials were ignored in the following analysis, this resulted in a final trial data set of size $N = 461$. The panels correspond to the values of χ and are ordered according to their average of the test responses. The horizontal axis is the rank of the trial determined by their test response. There is clear range of response values in the range $2 < \text{VAMP-2} < 4$ with the (ϕ, ψ, χ) dihedral angles (panel (a)) and the interatomic distances feature $(\lvert\mathbf{r}_1 - \mathbf{r}_2\rvert)$ (panel (c)) both performing the best out of the five features.

In contrast to the case of alanine dipeptide, all features show a marked degree of over-fitting Δf. However, within each feature there are combinations of the remaining hyperparameters for which $\Delta f = 0$ implying that it is possible to create a consistent picture of relaxation processes which generalize well for each feature. The response of the (ϕ, ψ, χ) feature approaches the maximum response of $\text{VAMP} - 2 = 4$ for the highest ranked trials, indicating the possibility of at least three slow relaxation processes.

The response surface as a function of the feature, χ, TICA lag time, τ, number of TICA components retains m, and the number of cluster centres, n was estimated with a GP. The RMSD

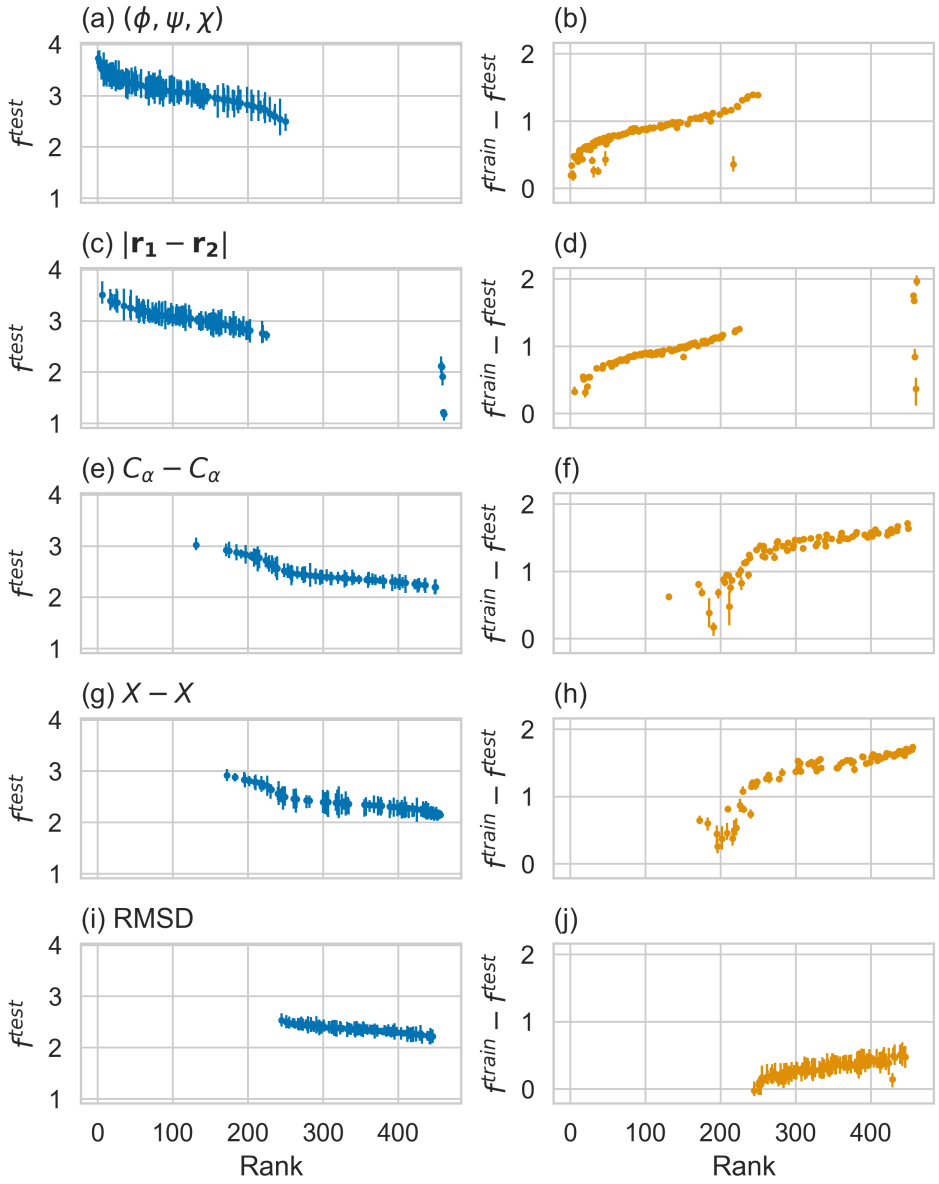


FIGURE 6.14: VAMP-2 SCORES OF THE HYPERPARAMETER TRIALS FOR MSMS OF AADH. The test response, $f^{\text{test}}(\mathbf{x};\mathbf{X}^{\text{test}})$, is shown in blue, for (a) the (ϕ,ψ,χ) dihedral angles, (c) heavy atom interatomic distances, (e) α-carbon contact distances, (g) heavy atom contact distances, (i) root mean square deviation of heavy atoms. The difference between f^{train} and f^{test} is shown in orange for (b) the the (ϕ,ψ,χ) dihedral angles and so on. The horizontal axis is the rank of the trial according to the test score. Each trial was scored with 20 iterations of 50:50 shuffle split cross-validation. The error bars represent the 25th and 75th quantiles of the cross-validation folds. The features are ordered according to the mean of the their test scores.

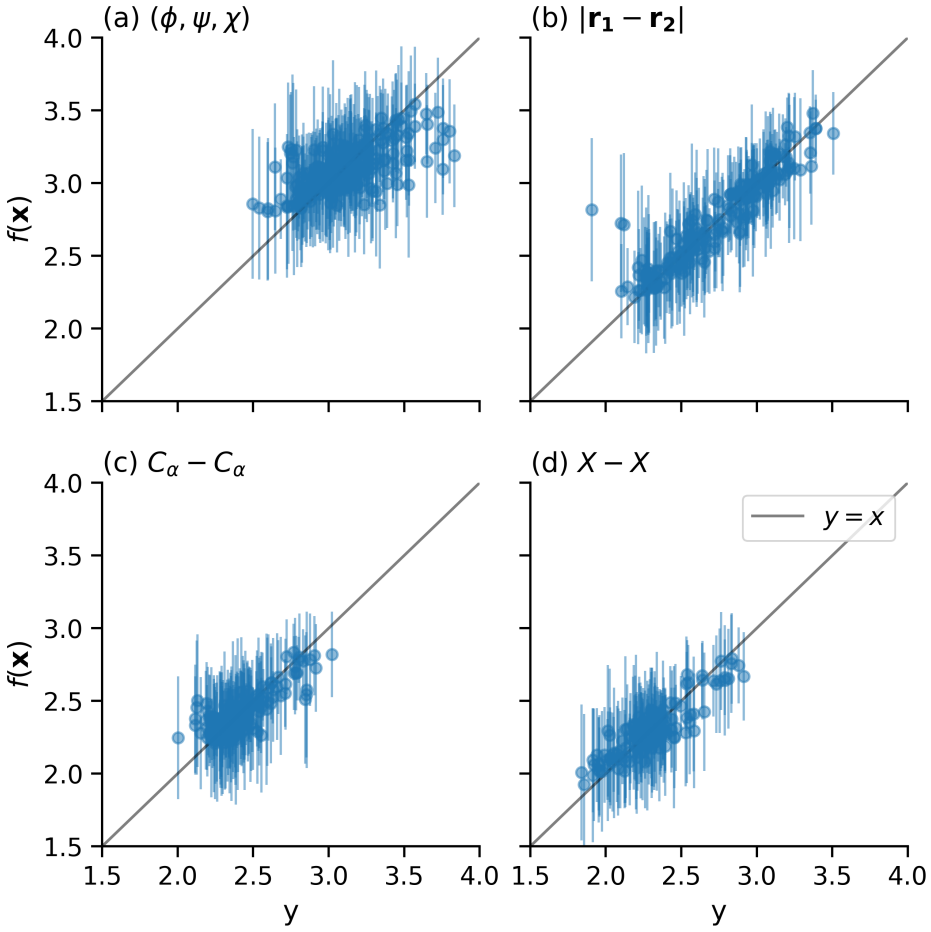


FIGURE 6.15: GOODNESS-OF-FIT FOR THE GP MODELLED RESPONSE SURFACE OF AADH. Panels (a) - (d) shows the goodness-of-fit conditional on each feature. The GP uses an exponential kernel with linear input warping on τ, m and n. The MSLL and SMSE were -0.01298 and 0.3087 respectively. The horizontal axis (y) are the observed trial values and the vertical axis is the mean response $(f(\mathbf{x}))$. The error bars are $\pm 2\sigma$ where σ is the standard deviation of the response surface including the noise term σ_n.

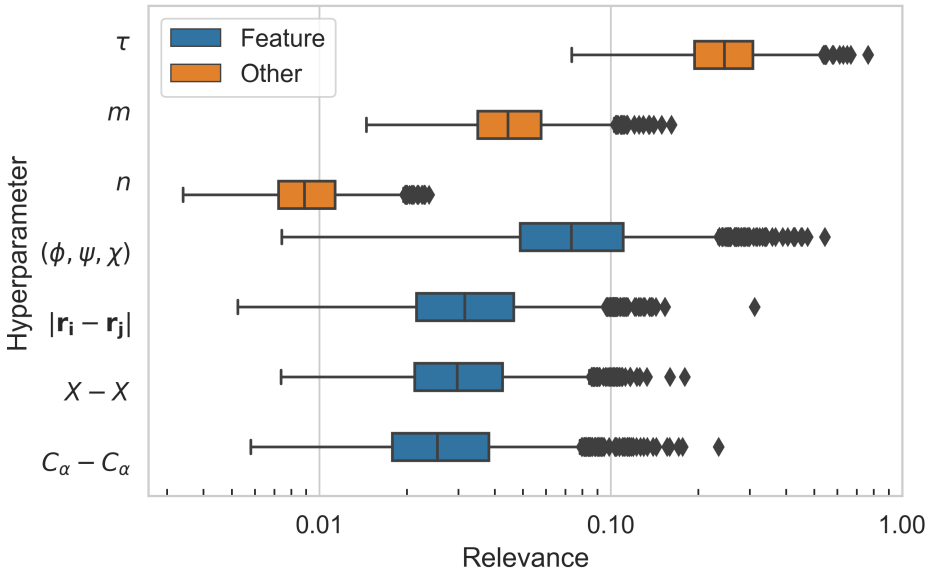


FIGURE 6.16: RELEVANCE OF THE HYPERPARAMETERS OF AADH. The distribution of the parameters of the response surface were estimated using MCMC with 1000 posterior samples. The relevance of the features (levels of χ) are shown in blue, labelled 'Feature'. The relevance of the the other hyperparameters are shown in orange, labelled 'Other'.

feature trials were removed from the data set because the τ and m hyperparameters do not apply, which would have created a *conditional* search space [135] which ca not easily be modelled by GPs (however, there are attempts to mitigate this [402]). While this is a problem in general for this method, the poor performance of RMSD as a feature both here and with alanine dipeptide means that this will not affect finding the optimum hyperparameters. There were also a number of trials which failed to converge an MSM, these were also removed. The final hyperparameter trial data set had $N = 361$ observations, $\mathscr{D}_{361}$. The covariance kernel was chosen from amongst the same set as for alanine dipeptide: exponential, Matérn 3-2, Matérn 5-2 and Gaussian (equations 4.14 - 4.17). In addition, the choice of logarithmic input warping was optionally applied to τ, m and n. The best combination, based on their combined rank using the MSLL and SMSE metrics, was an exponential kernel and a linear input transformation for τ, m and n. See table **??** for all the models' selection metrics. A more intuitive assessment of the fit of the can be found in figure 6.15 which shows the correlation between observed and predicted values for each feature. There is clearly a good fit for all the features except for (ϕ, ψ, χ) where the predicted values are slightly under and over estimated for the highest and lowest values with this feature respectively. This creates the possibility of a false bimodal response surface which must be checked when determining the optimal hyperparameters. The relatively poor fit on this feature is likely due to the fully multiplicative nature of the kernel. More flexible kernels (as discussed in e.g., reference [350]) which model lower order interactions may be able to overcome this problem in future work.

Hyperparameter	Median	(95 %C.I.)
$R_{(\phi,\psi,\chi)}$	0.073	(0.021-0.248)
$R_{\lvert\mathbf{r_1}-\mathbf{r_2}\rvert}$	0.032	(0.012-0.098)
$R_{C_\alpha - C_\alpha}$	0.025	(0.009-0.084)
R_{X-X}	0.030	(0.012-0.083)
R_τ	0.246	(0.126-0.474)
R_m	0.044	(0.022-0.095)
R_n	0.009	(0.005-0.018)
η	1.540	(1.117-2.154)
σ_n	0.011	(0.001-0.035)

TABLE 6.2: POSTERIOR DISTRIBUTIONS OF GP PARAMETERS. Shown are the median and 95% credible intervals for the kernel hyperparameters of the AADH response surface estimated using MCMC with 1000 posterior samples. The length-scale parameters in equation 4.13 are re-written here as relevances.

The multidimensional nature of the response surface poses problems for visualisation and for understanding the interaction between the hyperparameters in determining the response. However, calculating the hyperparameter relevance can help by suggesting the displayed granularity of the inputs. Figure 6.16 shows the posterior distribution of relevance for the features (blue) and the remaining hyperparameters (orange). The median and 95% credible intervals are tabulated in table 6.2.

Figure 6.17 shows a projection of the response surface, $f(\mathbf{x})$, informed by the relevance. τ and m are the two highest relevance hyperparameters ($R_\tau = 0.246$ [0.126–0.474)], $R_m = 0.044$ [0.022–0.095]) so the response surface is shown as a 2D heat map with τ on the vertical and m on the horizontal axis. Only odd values of m are shown given the slightly lower relevance of this feature. The number of cluster features is, like the case of alanine dipeptide, the lowest relevance hyperparameter ($R_n = 0.009$ [0.005–0.018] and so heat maps for only two value of n are shown: the value at the maximum of the response surface ($n = 207$ although the value displayed is rounded to 210) and $n = 1000$. With only four features, it is simple to show the response surface for each of them. With a larger number of features, displaying the high relevance features and *only one* of the low relevance features would be sufficient. This is because low relevance features are similar. Note, that this does not mean they have a low absolute value but rather that their values are highly correlated. This is clearly borne out with this surface - the two lower relevance features, the contact distances, are very similar and including both in the visualisation is redundant. However, as all features have absolutely low relevance ($R < 1$) then their response surfaces are expected to be similar. The maximum of the response surface, shown highlighted with a white star, occurs at $\mathbf{x} = \left(\chi = (\phi,\psi,\chi), \tau = 12.5\,\mathrm{ns}, m = 1, n = 207\right)$ with a value of $\mu = 3.543(396)$. The features of this response surface will be discussed in the context of sensitivity analysis in section 6.3.4.

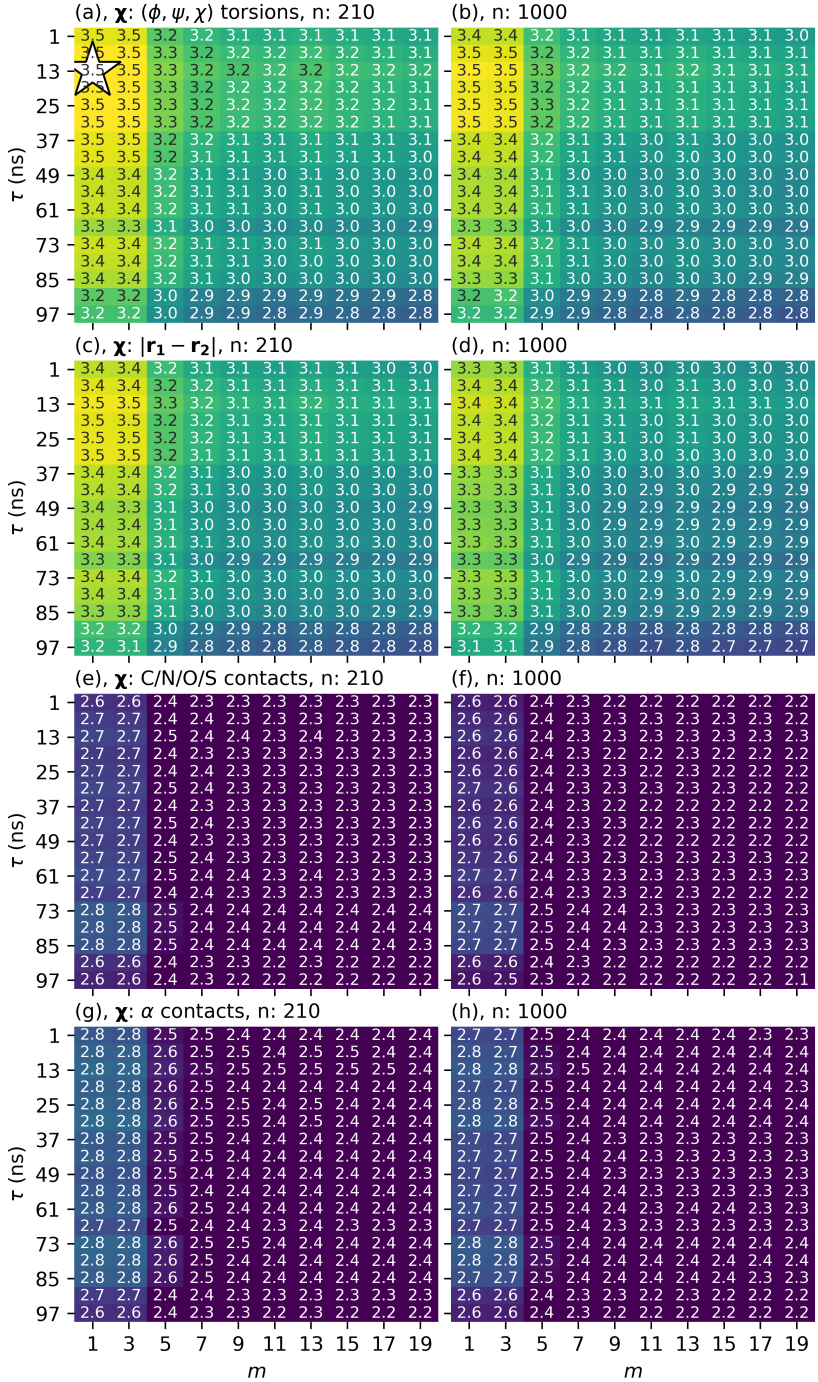


FIGURE 6.17: UNOPTIMIZED RESPONSE SURFACE OF AADH, $f(\mathbf{x};\mathscr{D}_{361})$. For each feature two heat maps are shown with τ on the vertical axis and m on the horizontal axis. Panel (a) shows the (ϕ,ψ,χ) feature with $n=210$, panel (b) for $n=1000$, and similarly for the remaining features. The white star denotes the approximate location of the maximum of the surface, the true maximum occurs at $\tau=12.5\,\text{ns}, m=1, n=207$ The value of the response surface denoted by the color (lighter implies higher values) and with the text annotations.

(Hyper)parameter	N_{seed}	100					361
	#	1	2	3	4	5	1
N_{total}	Pre	100	100	100	100	100	361
	Post	150	150	150	150	150	410
μ	Pre	3.487	3.355	3.599	3.473	3.478	3.543
	Post	3.500	3.558	3.545	3.581	3.569	3.558
σ	Pre	0.152	0.135	0.306	0.234	0.126	0.198
	Post	0.202	0.117	0.101	0.084	0.072	0.091
χ	Pre	(ϕ,ψ,χ)	$\lvert\mathbf{r_i}-\mathbf{r_j}\rvert$	(ϕ,ψ,χ)	(ϕ,ψ,χ)	$\lvert\mathbf{r_i}-\mathbf{r_j}\rvert$	(ϕ,ψ,χ)
	Post	(ϕ,ψ,χ)	(ϕ,ψ,χ)	(ϕ,ψ,χ)	(ϕ,ψ,χ)	(ϕ,ψ,χ)	(ϕ,ψ,χ)
τ	Pre	51.0	57.8	12.5	12.5	18.0	12.5
	Post	1.0	4.0	3.0	1.0	1.0	10.0
m	Pre	1	3	1	1	1	1
	Post	2	1	2	1	1	2
n	Pre	396	234	207	207	79	207
	Post	10	180	10	230	540	310

TABLE 6.3: MSM HYPERPARAMETERS FOR AADH PRE- AND POST-BAYESIAN OPTIMISATION. Each column represents a Bayesian optimisation experiment, seeded with N_{seed} randomly sampled hyperparameter trials. Five iterations of optimisation were run with $N_{\text{seed}} = 100$ (labelled #1,2 etc.) and a single iteration optimising the response surface using all the trial data ($N_{\text{seed}} = 361$). Each row is a variable or outcome with values associated with the optimum value of μ before and after Bayesian optimisation.

6.3.3 Optimisation

In order to test the convergence of the maximum of $f(\mathbf{x};\mathscr{D}_{361})$, 50 steps of Bayesian optimisation was performed. The optimisation was seeded with all the hyperparameter trial observations using the GP model determined in section 6.3.2. At each step of the optimisation, candidate MSM hyperparameters were determined as those which had the highest expected improvement. The grid of points, $\mathbf{X}_M$, used to maximize the acquisition function was a $4\times100\times20\times100$ ($\chi\times\tau\times m\times n$) evenly spaced grid. The values of the response and associated hyperparameters pre- and post-optimisation (i.e. $f(\mathbf{x};\mathscr{D}_{361})$ and $f(\mathbf{x};\mathscr{D}_{410})$) are tabulated in the final column of table 6.3. The optimisation resulted in a small improvement in the response from $\mu = 3.543\pm0.396 \rightarrow 3.558\pm0.182$ (a 0.4% improvement) with no change in the optimum feature (ϕ,ψ,χ) and only small changes in the other hyperparameters. The final optimised hyperparameters were: $\chi = (\phi,\psi,\chi)$, $\tau = 10\,\text{ns}$, $m = 2$, $n = 310$.

In order to see whether this maximum could be reached with a fewer number of trials, five response surfaces fit on random subsets of the trial data with 100 observations, $f^i(\mathbf{x};\mathscr{D}_{100})$, $i = 1-5$, were optimized with 50 steps of Bayesian optimisation. The kernel and input warping for $f^i(\mathbf{x};\mathscr{D}_{100})$ were determined for each i separately. The model selection metrics for each combination of kernel and input warping are tabulated in tables 11.2 - 11.6. The results of the optimisation trials are

shown in table 6.3 and the incumbent trajectories are shown in are shown in figure 6.18. Also shown, for comparison, in figure 6.18 are the maximum of the response surface $f(\mathbf{x};\mathscr{D}_{361})$ as orange lines, and the maximum of $f(\mathbf{x};\mathscr{D}_{410})$ as blue lines.

From table 6.3 it can be seen that after Bayesian optimisation each subset had maxima indistinguishable from the maxima after seeding with the full data set $f(\mathbf{x};\mathscr{D}_{461})$: each had the same optimal feature ((ϕ,ψ,χ) dihedral angles), slightly smaller values of τ ($1\,\mathrm{ns}$ to $4\,\mathrm{ns}$ cf. $10\,\mathrm{ns}$), similar values of m (1–2 cf. 2) but with a range of different values of n (10–540 cf. 310). The extent to which these differences in response surface maxima make a significant difference the final MSM, will be discussed in section 6.3.4, in the context of the sensitivity analyses.

It is clear that Bayesian optimisation only had a small effect on the value of the incumbent and on the optimum hyperparameters. The value of incumbent throughout the procedure (figure 6.18 panels (a) - (e)) remained relatively constant. The largest increase came from iteration 4, with $\Delta\mu = 0.108$ and there was even a slight decrease in iteration 3 with $\Delta\mu = -0.054$ although when including uncertainty they were statistically indistinguishable from each other. The optimisation procedure also had negligible effects on the value of χ, m and n. It did however, explore large values of τ before settling on its final optimum value in most of the iterations.

The optimisation procedure did not strongly increase the incumbent of each iteration, however, the incumbent and optimum hyperparameters of $f^i(\mathbf{x};\mathscr{D}_{150})$ are almost indistinguishable from $f(\mathbf{x};\mathscr{D}_{461})$. This suggests that the Bayesian optimisation procedure could be seeded with fewer than 100 trials if a GP model of the response surface could be estimated reliably.

6.3.4 Sensitivity analysis

So far the response surface for AADH has been modelled and optimised with a combination of random hyperparameter sampling and Bayesian optimisation under two assumptions about the model specification: that $\tau_{\mathrm{M}} = 2\,\mathrm{ns}$ and $r = 4$ are reasonable assumptions about the lag time and number of slow relaxation processes. As these were established on the basis of the reference MSM it is important to test these assumptions are still reasonable with the final optimized MSM.

To test the suitability of τ_{M} the implied timescales of a series of MSMs fit with the optimum hyperparameters but different lag times are shown in figure 6.19. Panel (a) shows shows the implied timescales for $\tau_{\mathrm{M}} = 0.1\,\mathrm{ns}$–$5\,\mathrm{ns}$ which clearly shows there is a slight increase in the top two implied timescales t_2 and t_3 suggesting that τ_{M} may be too small. Looking at the implied timescales over a larger range, $\tau_{\mathrm{M}} = 0.1\,\mathrm{ns}$–$50\,\mathrm{ns}$ (panel (b)), shows that there is a small plateau in t_2 and t_3 around $15\,\mathrm{ns}$ to $20\,\mathrm{ns}$. It is therefore not possible on this evidence alone to say definitively whether τ_{M} should be 2 or $20\,\mathrm{ns}$ or whether it makes a difference to non-quantitative aspects of the final model.

At $\tau = 2\,\mathrm{ns}$ there is no clear separation of timescales between the first four relaxation processes (taking into account the uncertainty in t_2 and t_3), however, there is a gap between t_5 and t_6 suggesting $r = 5$ would be more appropriate than the $r = 4$ previously chosen. This means the fourth relaxation process (t_5, shown in red in figure 6.19) has not been factored into the response.

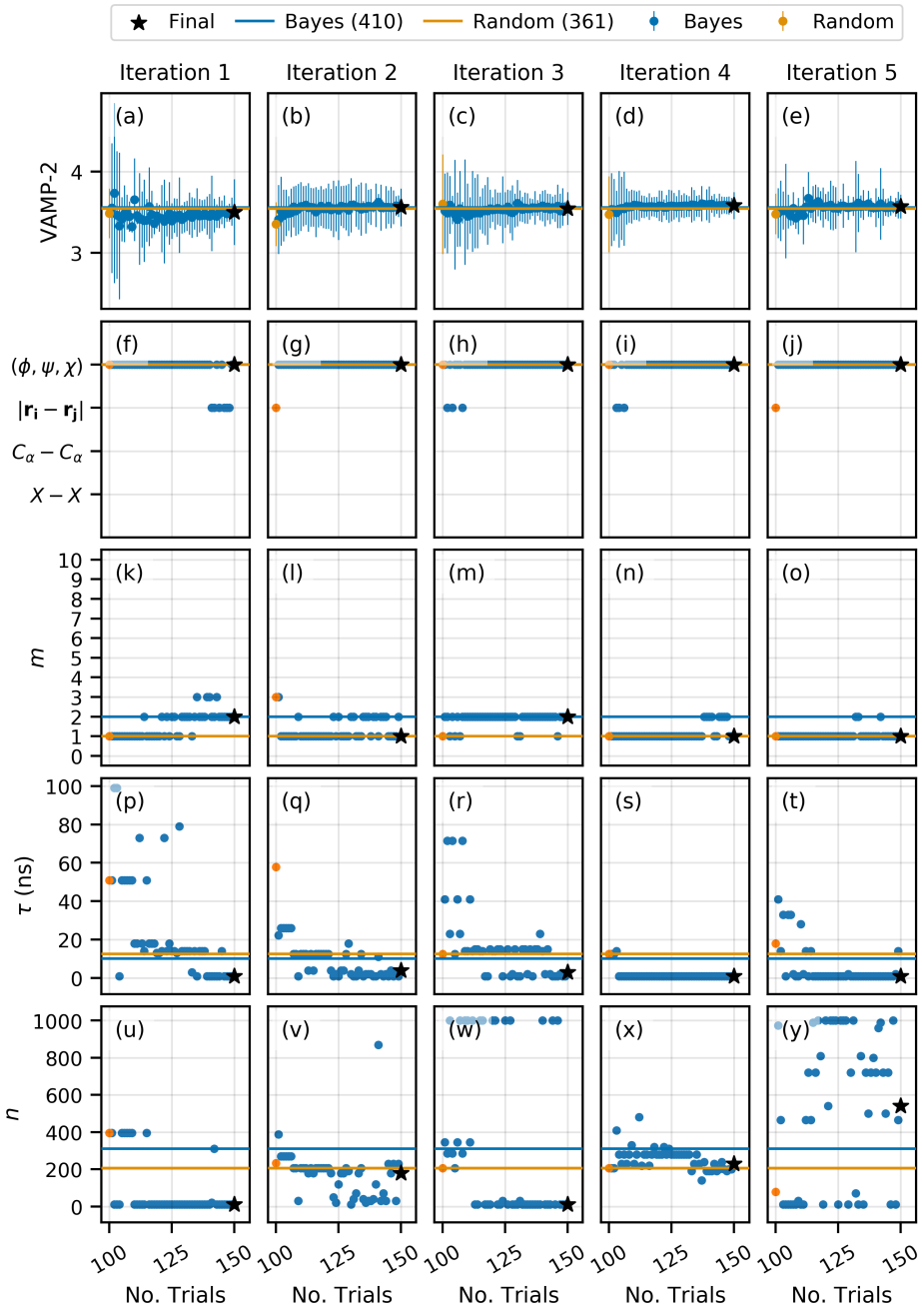


FIGURE 6.18: BAYESIAN OPTIMISATION TRAJECTORIES. Each column shows the trajectories of the incumbent $\mu^* \pm 2\sigma$ (panels (a) - (e)), and the accompanying value of: χ (panels (f) - (j)), m (panels (k) - (o)), τ (panels (p) - (t)) and n (panels (u) - (y)). The location of the final value of the trajectory is highlighted with a black star. For reference the corresponding values determined from the response surface using all the hyperparameter trials are shown as solid horizontal lines: the response surface before optimisation ($f(\mathbf{x};\mathscr{D}_{361})$) are shown in orange and after Bayesian optimisation ($f(\mathbf{x};\mathscr{D}_{410})$) in blue.

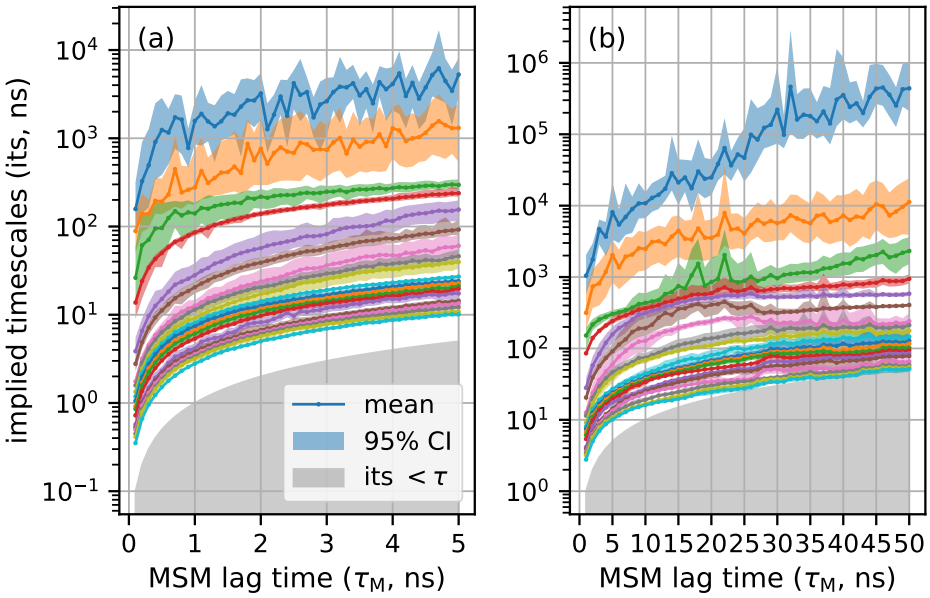


FIGURE 6.19: IMPLIED TIMESCALES OF THE BASE CASE MSM. Timescales were estimated using MCMC with 500 posterior samples. Panel (a) shows the implied timescales for $\tau_M = 0.1\,\text{ns}$–$5\,\text{ns}$ and panel (b) for $\tau_M = 0.1\,\text{ns}$–$50\,\text{ns}$. The solid lines are the mean of the posteriors, the shaded areas are the 95 % credible intervals. The grey shaded area is the region for which the implied timescales are smaller than τ_M.

In principle this means that there may be different hyperparameters which maximize VAMP-2, however, given the similarity between t_4 and t_5 this may not be very plausible or significant, if true. At $\tau_M = 20\,\text{ns}$ there is a clear separation between t_2 and t_3 suggesting $r = 2$ is appropriate. In this case we have included potentially too many fast processes. While changing the value of r will certainly change the value of the optimal hyperparameters, given the mixed evidence in the eigenvalue spectrum there is no reason to reject this optimized MSM. The associated question of how many metastable states and dominant relaxation processes actually matter for explaining the observed data will be investigated in section 6.4.

Knowledge of the response surface and of the eigenvalue spectrum suggests sensitivity analyses to understand the validity and robustness of the optimum MSM. The goal of sensitivity analyses is to have faith that reasonable changes in model choices and hyperparameters do not materially affect inferences from the model. Typically we are concerned with inferring relaxation timescales (t_i), the character of the relaxation process (Ψ_i) and the lumping of the microstates into metastable states. The VAMP-2 score has served as a proxy for the quality of the inferences required from the model but this is not sufficient for a number of reasons. First, in general it may be sensitive to both the MSM lag time and the number of eigenfunctions used in the definition. Second, as figure 4.10 has demonstrated, VAMP-2 is not sensitive to the discretisation error. This is also evident in the fact that in both iteration 1 and 3 (table 6.3) the optimal value of $n = 10$ with only a very small difference in response at these values $\mu = 3.500$ and $\mu = 3.545$ cf. $\mu = 3.558$). Third, the phenomenon

of the Rashomon effect [318] in statistical modelling, where multiple *different* statistical models result in the performance metric, could be at play here.

The standard validation check of MSMs, the Chapman-Kolmogorov test (section 2.5), relies on coarse graining a transition matrix, which will be discussed in section 6.4. The current discussion will center on the eigenvalue spectrum and qualitative aspects of the free energy surface and eigenfunctions. For the optimal MSM (hereafter labelled 'base case'), these are all shown in figure 6.21: panels (a) - (c) show the eigenvectors of the first three relaxation processes ($\Psi_2 - \Psi_4$) in the space of the first two TICA components. Panel (d) shows the implied timescales for the first 10 relaxation processes (coloured according to whether they were included in the VAMP-2 score), and panel (e) shows the free energy surface in the space of the first two TICA components.

Figure 6.21 will serve as a base case for a number of sensitivity analyses, informed by the optimised response surface (figure 6.20) and eigenvalue spectrum (figure 6.19). A summary of the hyperparameters for the base case and sensitivity cases are shown at the end of this section in table 6.4.

Sensitivity 1 changed the MSM lag time from $2\,\mathrm{ns}$ to $20\,\mathrm{ns}$ and is shown in figure 6.22. As expected, the absolute values of the implied timescales have increased, as has the relative separation between t_2 and t_3. This sensitivity case is included as, while the sign structure is the same for the first and third relaxation process, it has changed for the second relaxation process. Therefore the first, definitely slow, relaxation process is robust with respect to its sign structure. The second relaxation process which may or may not be dominant, changes its character as a function of the MSM lag time and will need to be further investigated.

Sensitivity 2 changed the hyperparameters to the best performing set with the interatomic distances feature ($\chi = |\mathbf{r}_1 - \mathbf{r}_2|,\ \tau = 1\,\mathrm{ns},\ m = 2,\ n = 110$) and is shown in figure 6.23. This was justified because of the similarity in the maximum response values (base case: $\mu = 3.56 \pm 0.18$, sensitivity 2: $\mu = 3.44 \pm 0.35$). There is a clear similarity between sensitivity 2 and the base case in both the free energy surface and the first relaxation processes' timescale ($t_2 = 2.7\,\mu\mathrm{s}$ cf. $t_2 = 2.2\,\mu\mathrm{s}$ in the base case) and sign structure of Ψ_2 and Ψ_3. There is also a greater separation in timescales between t_2 and t_3 ($t_2/t_3 \simeq 6$) than the base case ($t_2/t_3 \simeq 2$). Taken together these two observations suggest that the interatomic distances potentially resolve the first relaxation processes similarly.

Sensitivity 3, changed the value of τ to $85\,\mathrm{ns}$ from $\tau = 10\,\mathrm{ns}$ in the base case and is shown in figure 6.24. This value of τ was chosen because the response at this point had the smallest overlap with the incumbent (base case: $\mu = 3.56 \pm 0.18$, sensitivity 3: $\mu = 3.30 \pm 0.24$), in other words it was the least similar of the values with overlapping distributions. Here there is a distinct difference in the absolute values of the timescales and the free energy surface compared to the base case. To further delineate the difference between the two TICA representations, figure 6.25 shows the difference between the first two TICA components. The TICA components of the base case are shown in blue with the magnitude of the individual elements (corresponding to the 116 dihedral angle features) ordered in decreasing value. The corresponding elements for sensitivity 3 are shown

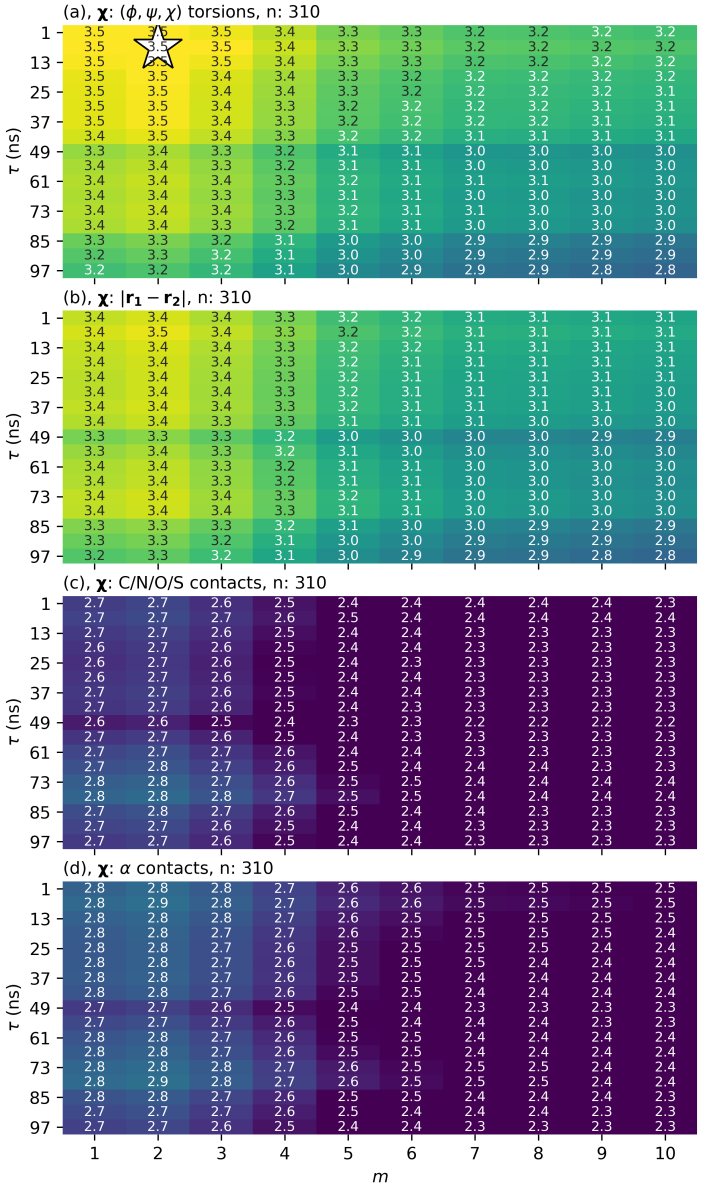


FIGURE 6.20: OPTIMISED RESPONSE SURFACE OF AADH, $f(\mathbf{x};\mathscr{D}_{410})$. The response surface was estimated using both the randomly sampled and optimised hyperparameters. For each feature a single heat map is shown for $n = 310$, with τ on the vertical axis and m on the horizontal axis. All values of m in the range $1 \leq m \leq 10$ are shown. The white star denotes the approximate location of the maximum of the surface, the true maximum occurs at $\tau = 10\,\text{ns}, m = 2, n = 310$. The value of the response surface is denoted by the color (lighter implies higher values) and with the text annotations.

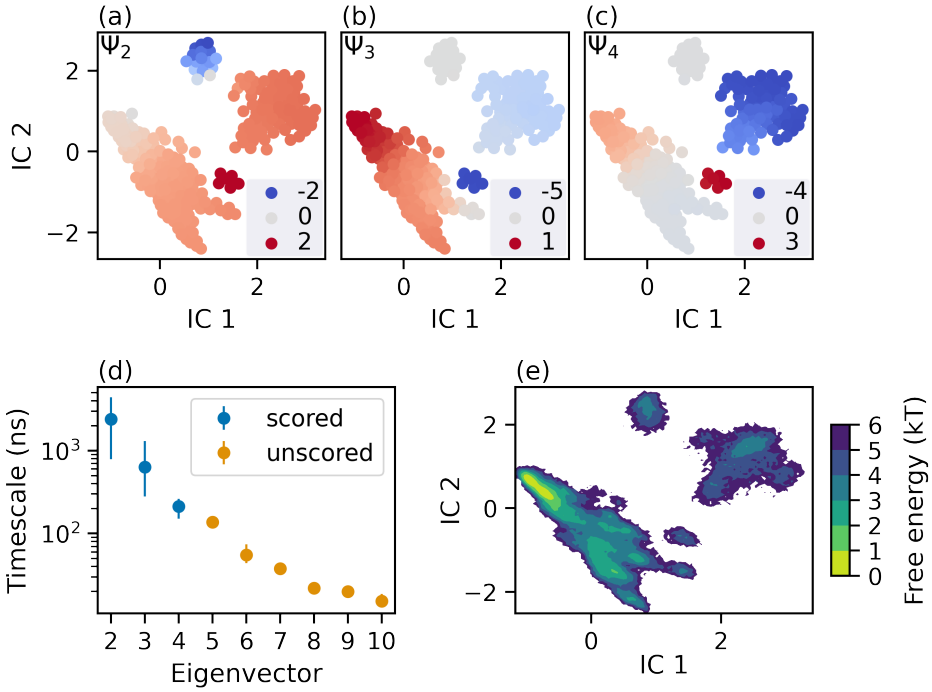


FIGURE 6.21: BASE CASE MSM. The MSM was estimated with the optimum hyper-parameters: $\chi = (\phi, \psi, \chi)$, $\tau = 10\,\text{ns}$, $m = 2$ and $n = 310$. Panels (a), (b) and (c) show the non-trivial eigenvectors used in the VAMP-2 score, the horizontal and vertical axes are the first two TICA components. Panel (d) are the first ten implied timescales, colored according to whether they were used in the VAMP-2 score. The error bars are the 95% credible intervals. Panel (e) is the free energy with the same axes as panels (a) - (c). The MSM was estimated using MCMC with 1000 posterior samples.

in orange with the sign flipped. The normalized overlap between the two TICA components of the base case and sensitivity 3 are 0.63 and 0.56 for the first and second TICA components respectively. This is due to the different weights attached to each dihedral angle features. This suggest that the two TICA eigenvectors represent a qualitatively and quantitatively different model. Where this change comes in the response surface and and what the corresponding values of the response will determine whether this an example of the Roshomon effect.

6.4 Coarse grained model

The aim of this section is to produce a coarse-grained picture of the conformational dynamics using a hidden Markov model. For each of the Markov state models defined in table 6.4, maximum likelihood HMMs with between $2-20$ hidden states were estimated and the ICL calculated for each one. No striding of the data was performed. The number of hidden states was determined by the smallest value of the ICL (g^{ICL}). A Bayesian HMM was estimated with the selected number of

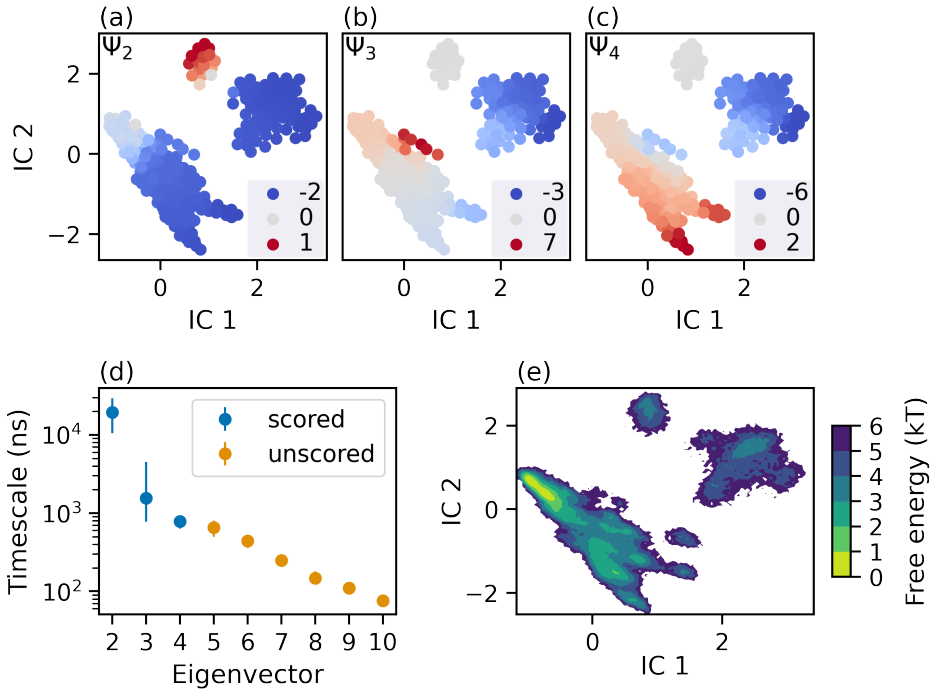


FIGURE 6.22: SENSITIVITY 1 MSM. The MSM has the same hyperparameters as the base case but with $\tau_{\mathrm{M}} = 20\,\mathrm{ns}$. See caption of figure 6.21 for details.

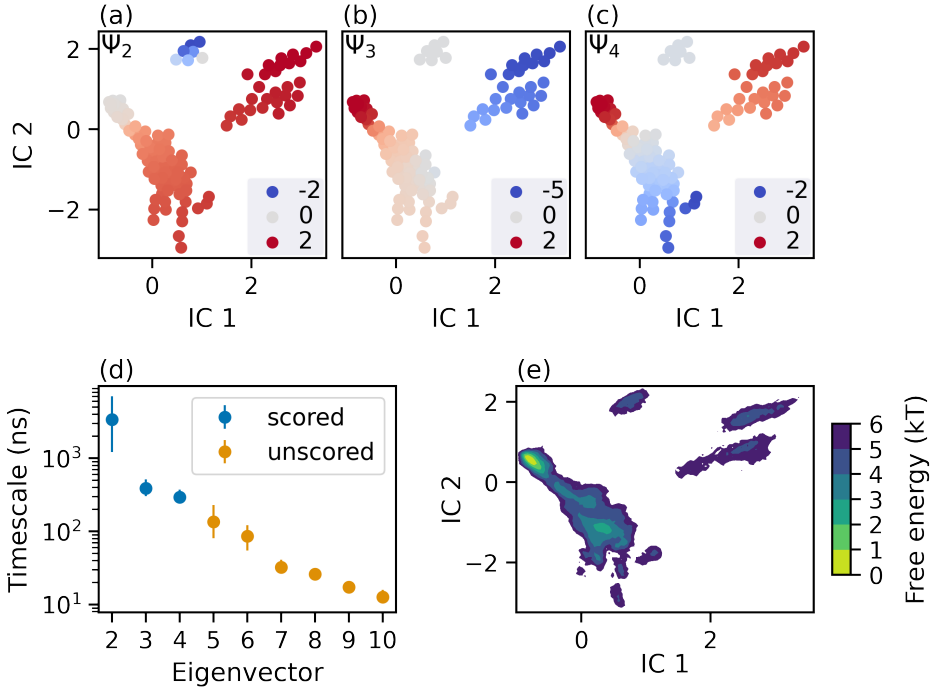


FIGURE 6.23: SENSITIVITY 2 MSM. This sensitivity used the best performing hyperparameters with a different value of χ to the base case. The hyperparameters were chosen to be $\chi = |\mathbf{r}_1 - \mathbf{r}_2|$, $\tau = 1\,\mathrm{ns}, m = 2, n = 110$) and $\tau_{\mathrm{M}} 2\,\mathrm{ns}$. See caption of figure 6.21 for details.

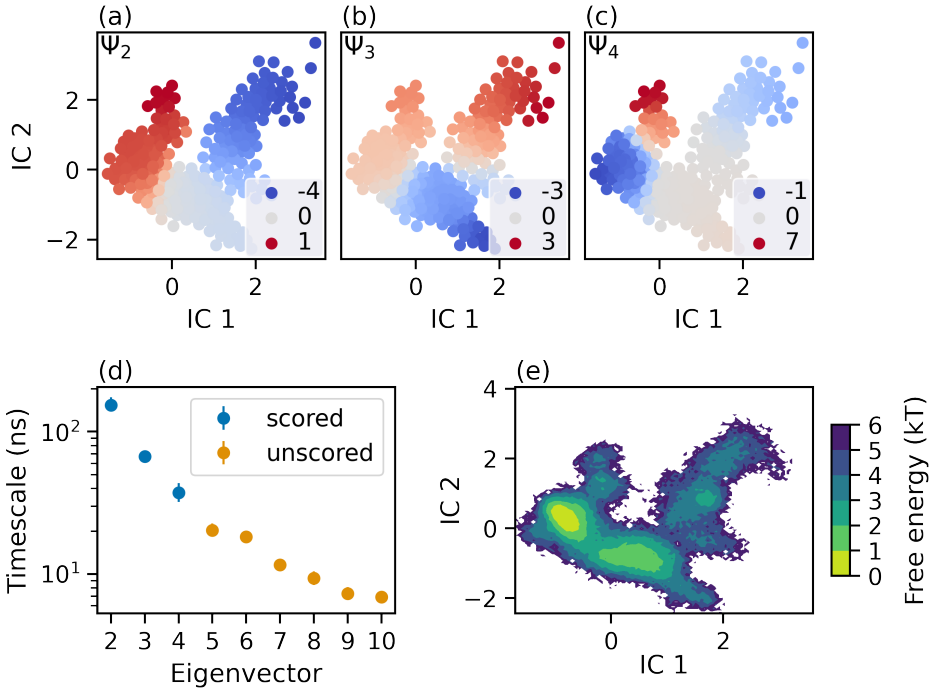


FIGURE 6.24: SENSITIVITY 3 MSM. The hyperparameters were the same as the base case but with a longer TICA lag time $\tau = 85\,\text{ns}$. See caption of figure 6.21 for details.

Parameter	Base case	Sensitivity 1	Sensitivity 2	Sensitivity 3
Markov lag time, τ(MSM)	2 ns	20 ns	2 ns	2 ns
Feature, χ	(ϕ, ψ, χ)	(ϕ, ψ, χ)	$\lvert\mathbf{r}_1 - \mathbf{r}_2\rvert$	(ϕ, ψ, χ)
TICA lag time, τ	10 ns	10 ns	1 ns	85 ns
TICA components, m	2	2	2	2
Cluster centres, n	310	310	110	310

TABLE 6.4: MARKOV LAG TIME AND HYPERPARAMETERS OF SELECTED MODELS. These models are the base case and sensitivity models to be coarse grained in section 6.4.

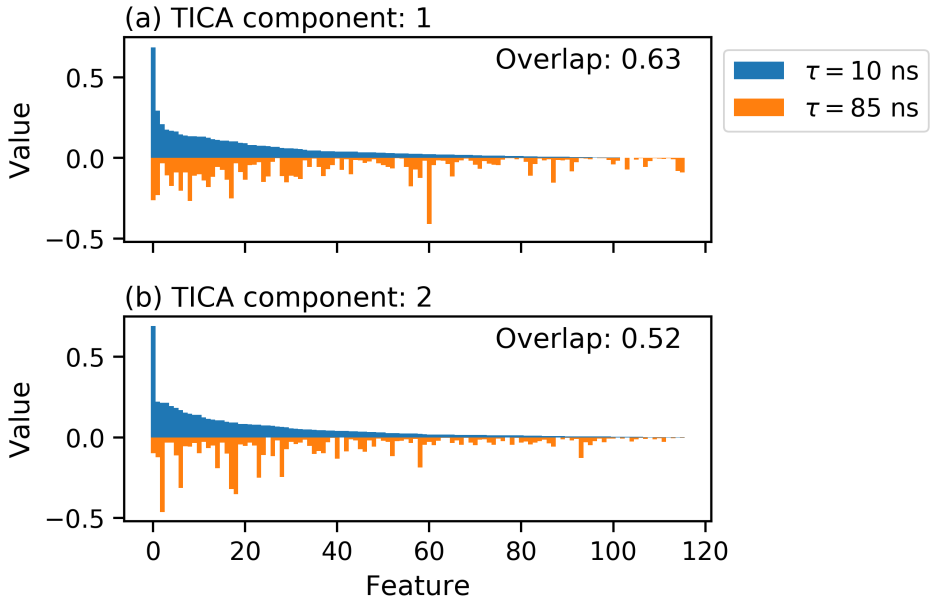


FIGURE 6.25: COMPARISON OF THE BASE CASE AND SENSITIVITY 3 TICA EIGENVECTORS. The TICA lag time for the base case was $\tau = 10\,\mathrm{ns}$ (blue) and for sensitivity 3 was $\tau = 85\,\mathrm{ns}$ (orange). The individual elements correspond to the weights associated with each of the 116 dihedral angle features. Only the absolute values are shown. The elements were ordered according to the absolute value of the elements in the base case for each TICA component to highlight the differences. The normalized overlap between the base case ($\mathbf{v}_1$) and sensitivity three ($\mathbf{v}_2$) is calculated as $\mathbf{v}_1 \cdot \mathbf{v}_2 / |\mathbf{v}_1||\mathbf{v}_2|$.

hidden states with four independent chains each with 4000 posterior samples, collected after 1000 burn-in steps. The trajectories were strided according to equation 2.31 to avoid optimistic bias in the error estimates [248]. Convergence of the samples was checked using the rank-normalized $\hat{R}$ statistic [254] of the non-zero hidden transition matrix elements. If a converged Bayesian HMM could be estimated then a CK test was performed. All calculations were performed using the packages cited in chapter 5.

Values of the ICL for each case are shown in figure 6.26 which have been scaled for clarity. For reference the unscaled ICL values and the contribution due to the classification entropy are tabulated in table 11.7. The ICL behaves differently for AADH compared to the Prinz potential and selects large numbers of hidden states. As expected the log-likelihood increases with g for each case, as does the BIC penalty term ($d \cdot \log(N_{\mathrm{obs}})$), however, unlike the results for the Prinz potential (figure 5.3) the entropy term increases only negligibly. In addition, as the maximum likelihood HMMs were calculated using the sliding window counting method, the value of N_{obs} is an over-estimate. The number of selected states for each case is large (18, 8, 19, 18 for the base case and sensitivities 1 - 3 respectively) and in each case the number of strongly connected hidden states, g^{s}, was less than the stipulated number of hidden states, g (table 11.7 also shows g and g^{s} for all models). For example,

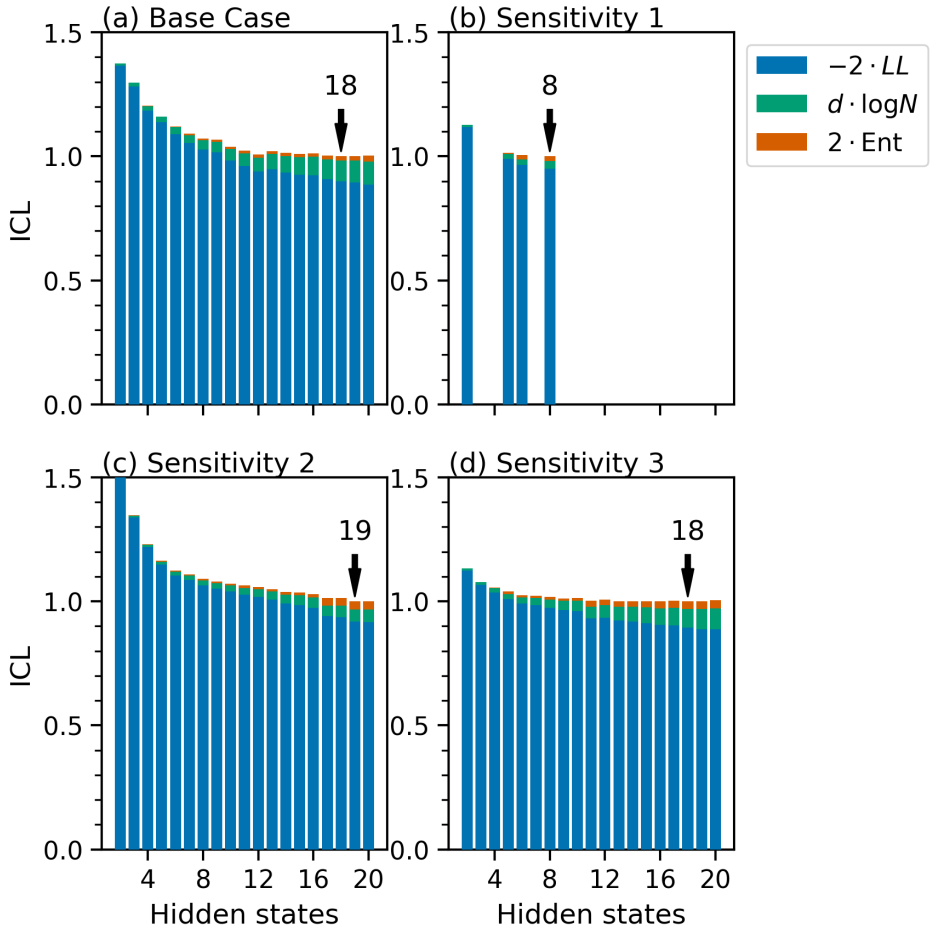


FIGURE 6.26: HIDDEN STATE SELECTION OF AADH WITH THE ICL. The base case is shown in panel (a), and the three sensitivity cases in panels (b) - (c). The coloured bars show the components of the ICL: the log-likelihood ($-2LL$) term is in blue, the BIC penalty term, $d \cdot \log N_{\mathrm{obs}}$, is shown in green, and the classification entropy penalty term, $2 \cdot EN$, is shown in red. The values are scaled so the minimum for each case is 1, this value is denoted with an arrow and labelled with number of hidden states. Models which failed to converge are missing.

the ICL selected $g^{\mathrm{ICL}} = 18$ hidden states for the base case, figure 6.26 panel (a). However, after re-estimating as a Bayesian HMM the largest strongly connected set within this model had $g^{\mathrm{s}} = 15$ hidden states. This was because, as described in section 2.4, the trajectories were first strided so that the observations were approximately independent. The striding interval was estimated from the data using equation 2.31 which for the base case was $\Delta t = 2\,\mathrm{ns}$. Increasing the value of τ_{M} will also decrease the connectivity of the count matrix if the hidden states are separated by timescales comparable to the new larger τ_{M}. This explains the failure to converge models 15 of the models at the longer lag time of sensitivity 1, figure 6.26 panel (b).

The base case with $g = 18$ hidden states was re-estimated using Bayesian estimation so that the errors could be determined. There was moderate convergence of the posterior chains with $63\,\%$ of the transition matrix elements under the recommended threshold of $\hat{R} < 1.01$ while the remaining elements had $\hat{R} < 1.1$ (the full $\hat{R}$ statistics are tabulated in table 11.8). The final model is shown in figure 6.27. Panel (a) shows the observed states assigned to the 15 strongly connected hidden states in the space of the two TICA components (IC 1 and IC 2). 267 out of 310 observed states are used in the model which constitute just under $95\,\%$ of the observed states found in the trajectories; the missing states are shown as unfilled black circles. However, the striding of the trajectories meant that only $5\,\%$ of the total observations were used. All metastable states have short lifetimes in the range $20\,\mathrm{ns}$ to $60\,\mathrm{ns}$ with two exceptions: state h_2 ($223\,\mathrm{ns}$ [$61.7\,\mathrm{ns}$–$4430\,\mathrm{ns}$]) and state h_{15}: ($275\,\mathrm{ns}$ [$53.4\,\mathrm{ns}$–$7400\,\mathrm{ns}$]). These two states are the two predominantly involved in the dominant relaxation process shown in panel (c) with a timescale of $1180\,\mathrm{ns}$ [$373\,\mathrm{ns}$–$8180\,\mathrm{ns}$]. This and the remaining timescales, for comparison, are shown in panel (d). There are two limitations to the base case hidden Markov model. First, the HMM could not be validated using the Chapman-Kolmogorov test because models with longer lag times resulted in further disconnections in the hidden state count matrix i.e., at lags of $\tau_{\mathrm{M}} > 2\,\mathrm{ns}$ the number of strongly connected hidden states was smaller than 15. Second, missing observed states around $(\mathrm{IC1} = 0.9, \mathrm{IC2} = 2.3)$ (which collectively will be called h_{-1}) in figure 6.27 panel (a) are involved in the the dominant relaxation process in the base case MSM, shown in figure 6.21 panel (a). This process, with an implied timescale of $2.21\,\mu\mathrm{s}$ [$1.03\,\mu\mathrm{s}$–$5.21\,\mu\mathrm{s}$] is therefore not captured in the HMM.

No Bayesian HMM with ICL selected number of hidden states could be estimated for either sensitivity 1 (the base case with $\tau_{\mathrm{M}} = 20\,\mathrm{ns}$) or for sensitivity 3 (the base case with $\tau = 85\,\mathrm{ns}$) because the trajectory striding required for accurate error estimation resulted in disconnected count matrix.

A Bayesian HMM was estimated for sensitivity 2 which, after striding by $2\,\mathrm{ns}$ according to equation 2.31, resulted in only 13 hidden states forming a strongly connected set. Convergence of the model was mixed, $60\,\%$ of the transition matrix elements were under the recommended threshold of $\hat{R} < 1.01$, however, the remaining elements were unconverged with values up to 1.3. The full list of $\hat{R}$ statistics can be found in table 11.11. The final model is shown in figure 6.28. Panel (a) shows the assignment of the observed states to the 13 hidden states. Only 89 of the 110 observed states were used in the model (the missing states are shown as unfilled circles) constituting

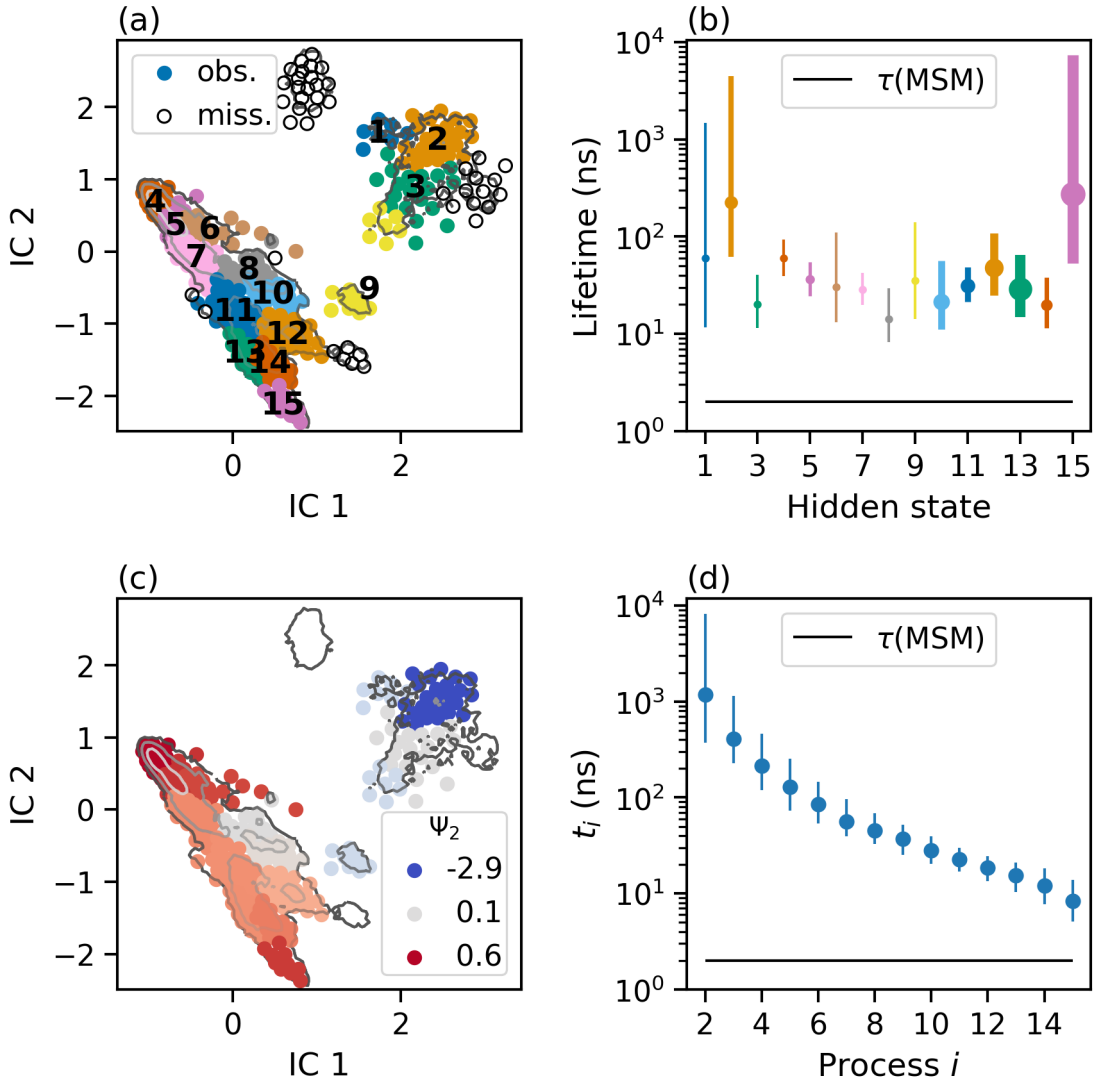


FIGURE 6.27: BASE CASE HMM. Panel (a) shows the observed states (colour circles) in the space of the first two TICA components, IC 1 and IC 2. Each observed state, j is assigned to the hidden state with the maximum a posteriori probability, i.e. $\mathrm{argmax}_i \mathrm{P}(h = i|s = j)$. Each colour corresponds to a hidden state, labelled 1–15. Panel (b) shows the median lifetimes of each state with 95 % credible intervals. The horizontal width of the error bars and sizes of the markers are proportional to the stationary distribution of each hidden state. The Markov lag time is shown as a black horizontal line for comparison. Panel (c) shows the dominant hidden state relaxation process, $\Psi_2(h)$. Each observed state assigned to a hidden state i is coloured according to the value of $\Psi_2(i)$. Panel (d) shows the median implied timescales of the hidden state relaxation processes with 95 % credible intervals. The Markov lag time is shown as a black horizontal line for comparison.

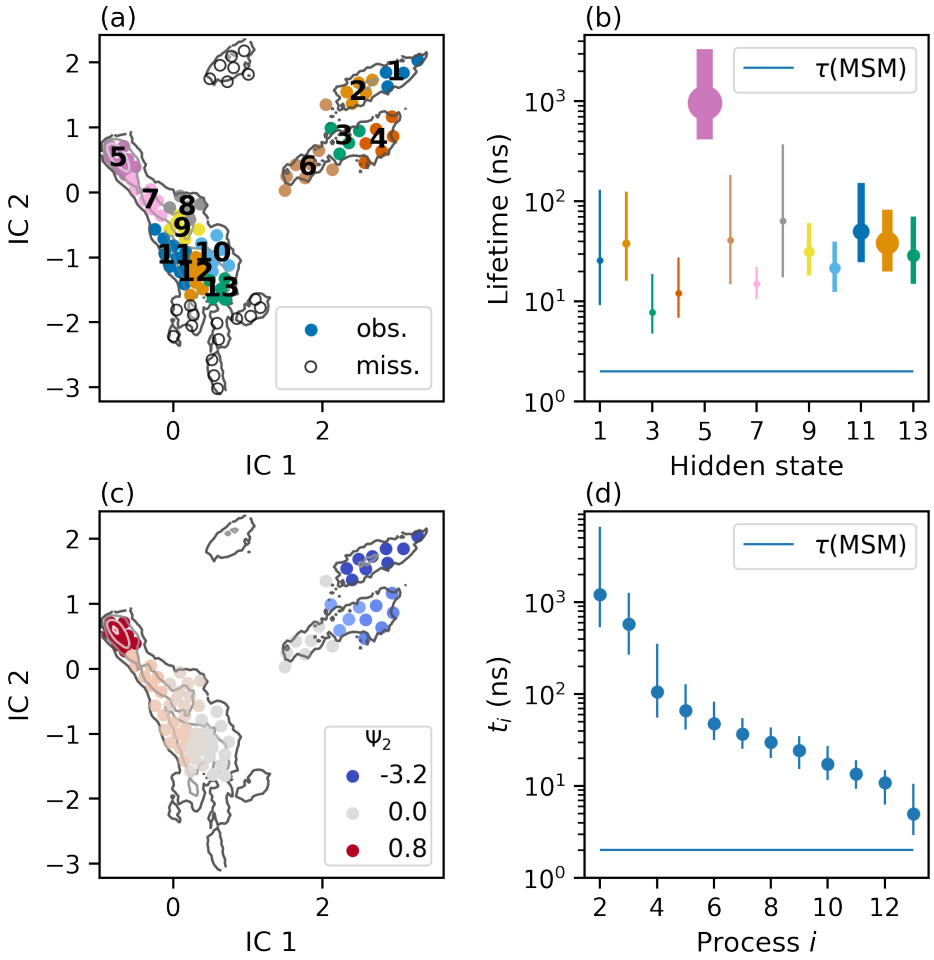


FIGURE 6.28: SENSITIVITY 2 HMM. Panel (a) shows the observed states (colour circles) in the space of the first two TICA components, IC 1 and IC 2. Each observed state, j is assigned to the hidden state with the maximum a posteriori probability, i.e. $\mathrm{argmax}_i \, \mathrm{P}(h = i | s = j)$. Each colour corresponds to a hidden state, labelled 1 to 13. Panel (b) shows the median lifetimes of each state with 95 % credible intervals. The horizontal width of the error bars and sizes of the markers are proportional to the stationary distribution of each hidden state. The Markov lag time is shown as a black horizontal line for comparison. Panel (c) shows the dominant hidden state relaxation process, $\Psi_2(i)$. Each observed state assigned to a hidden state i is coloured according to the value of $\Psi_2(i)$. Panel (d) shows the median implied timescales of the hidden state relaxation processes with 95 % credible intervals. The Markov lag time is shown as a black horizontal line for comparison.

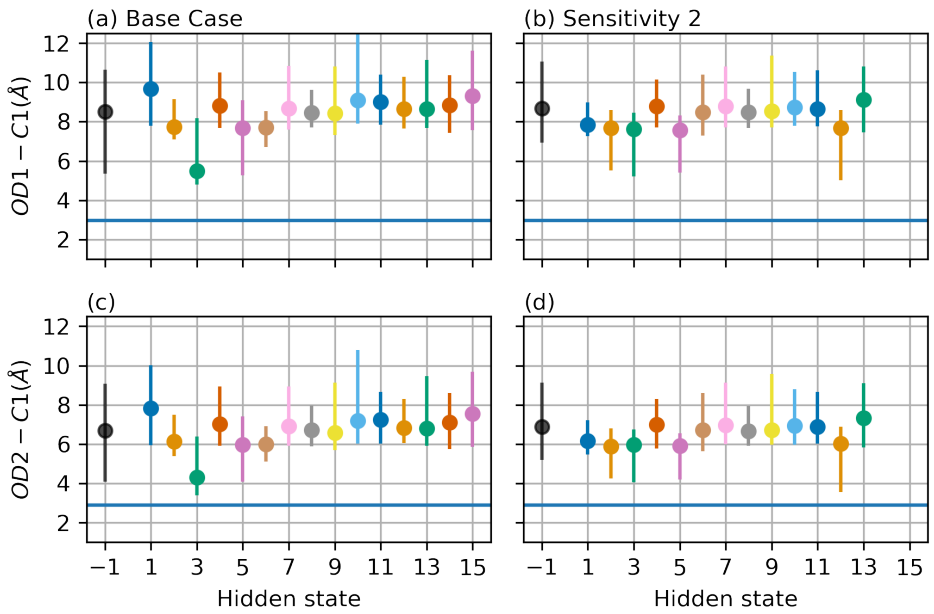


FIGURE 6.29: DONOR-ACCEPTOR DISTANCES OF THE HIDDEN STATES. The mean and 95 % quantiles of the distribution for the OD1—C1 and OD2—C1 for each hidden state in the base case (panels (a) and (c) respectively) and sensitivity 2 (panels (b) and (d) respectively) HMMs. The colour scheme matches the colours in figures 6.27 and 6.28. In addition, the missing states h_{-1} are shown in black. The horizontal blue line is distance estimated using QM/MM in reference [372].

over 92 % of the observed states. However, due to the striding only 4 % of the observations were used. The hidden state lifetimes are shown in panel (b). All hidden states were short-lived with median lifetimes between 8 ns to 64 ns except for state h_5, the most populous state, which has a lifetime of 961 ns [416 ns–3310 ns]. The dominant relaxation process, panel (c), involves population transfer between state h_5 and the four short-lived states $h_1 - h_4$ with an implied timescale of 1210 ns [533 ns–6590 ns]. This and the remaining implied timescales are shown in panel (d).

The sensitivity 2 HMM suffers from the same drawbacks as the the base case model and for the same reasons: the model could not be validated by the CK test and it is missing observed states important in the dominant relaxation process identified in the corresponding MSM, shown in figure 6.23. The collection of observed states in the region $(\mathrm{IC1} = 0.8, \mathrm{IC2} = 1.9)$ (which collectively will be called h_{-1}) in figure 6.28 panel (a) are involved in the the dominant process in the base case MSM, shown in figure 6.23 panel (a). This process, with an implied timescale of 2.69 µs [1.35 µs–4.61 µs is therefore not captured in the HMM.

6.5 Conformational landscape of AADH

Although the validity of the base case and sensitivity 2 models as a description of the dynamics has not been established, understanding their implications for the conformational dynamics will

complete their description and help guide further work in this area.

As already established in section 6.2.6 the sampled conformations in the D active site do not correspond to reactive conformations established in QM/MM studies. However, within the sampled conformations the base case HMM resolves hidden states with variable degrees of reactive character as determined by their acceptor-donor bond lengths (TTW-C1—Asp128-OD1/OD2 bond length) as shown in figure 6.29. The base case h_3 has a smaller donor-acceptor bond length to both both acceptors (OD1 and OD2), $5.5\,\text{Å}$ and $4.3\,\text{Å}$ respectively than the other hidden states. These other states (including the 'missing' state h_{-1}) have donor-acceptor lengths between $7.7\,\text{Å}$ to $9.7\,\text{Å}$ to OD1 and $6.0\,\text{Å}$ to $7.8\,\text{Å}$ to OD2.

The dominant relaxation process (figure 6.27 panel (a)) transfers population between the relatively populous h_2 ($\tilde{\pi}_2 = 10.3\,\%$ [0.1 %–73.4 %]) to state h_{15} ($\tilde{\pi}_{15} = 23.4\,\%$ [1.6 %–91.3 %]). From h_2 to h_{15} the donor-acceptor distance increases from approximately $7.7\,\text{Å}$ & $6.1\,\text{Å}$ to $9.3\,\text{Å}$ & $7.5\,\text{Å}$ to OD1 & OD2 respectively. However, this process proceeds through the 'reactive' state h_3 with the smallest pair donor-acceptor distances, $5.5\,\text{Å}$ & $4.3\,\text{Å}$.

This mechanism is shown in the network plot, figure 6.30. In this figure the hidden states i and j are only shown connected if the flux between them, $\tilde{\pi}_i \tilde{T}_{i,j}$, is greater than $0.01\,\%$. This is to simplify the picture - the full rate matrix is tabulated in table 11.9. The 'missing' state h_{-1} is also shown as a disconnected, dashed circle, for reference. Representative structures for a selection of the important states are also shown. In each of these structures bond to OD1, OD2, HI-2, HI-3, are highlighted in yellow and the three most important dihedral angles an highlighted in blue. These angles correspond to the top two largest TICA components in both IC 1 (χ_1-X109 and ϕ-G85, where X109 is the abbreviation of TTW109) and IC 2 (χ_1-X109 and ψ-V83).

In order to understand the slowest relaxation process in the base case HMM, consider the pathway $h_2 \to h_3 \to h_{13} \to h_{15}$. This is the highest flux pathway i.e. the pathway which at each point selects the highest flux connection. The change in the dihedral angles can be seen in the accompanying structures in figure 6.30 and in the distributions plotted in figure 6.31. The first step, $h_2 \to h_3$, proceed primarily through an increase in both ϕ-G85 (figure 6.31 panels (i) and (j)) and ψ-V83 (panels (e) and (f)). The second step, $h_3 \to h_{13}$, proceeds at a relatively slow rate of $0.8\,\mu\text{s}^{-1}$ [$0.0\,\mu\text{s}^{-1}$–$4.2\,\mu\text{s}^{-1}$] with a large increase in the χ_1-X109 angle (6.31 panels (b) and (c)) from approximately 70° to 190°. The final step h_{13} to h_{15} proceeds with a decrease in the variance of the ϕ-G85 and ψ-V83 distributions.

The picture arising from the coarse graining sensitivity 2 is very different to the base case, even though the difference in the estimated VAMP-2 score for the MSM was small, (base case: $\mu = 3.56 \pm 0.18$, sensitivity 2: $\mu = 3.44 \pm 0.35$). First, the hidden states of sensitivity 2 show no large differences in their donor-acceptor lengths which all lie between $7.6\,\text{Å}$–$9.1\,\text{Å}$ to OD1 and $5.9\,\text{Å}$–$7.3\,\text{Å}$ to OD2 as shown in figure 6.29 panels (b) and (d). The stationary distribution is distributed primarily in a single state, h_5 ($\tilde{\pi}_5 = 30.12\,\%$ [5.27 %–68.33 %]) which is an order of magnitude more long lived ($961\,\text{ns}$ [$416\,\text{ns}$–$3310\,\text{ns}$]) than either base case h_2 ($223\,\text{ns}$ [$61.7\,\text{ns}$–$4430\,\text{ns}$]) or h_{15}

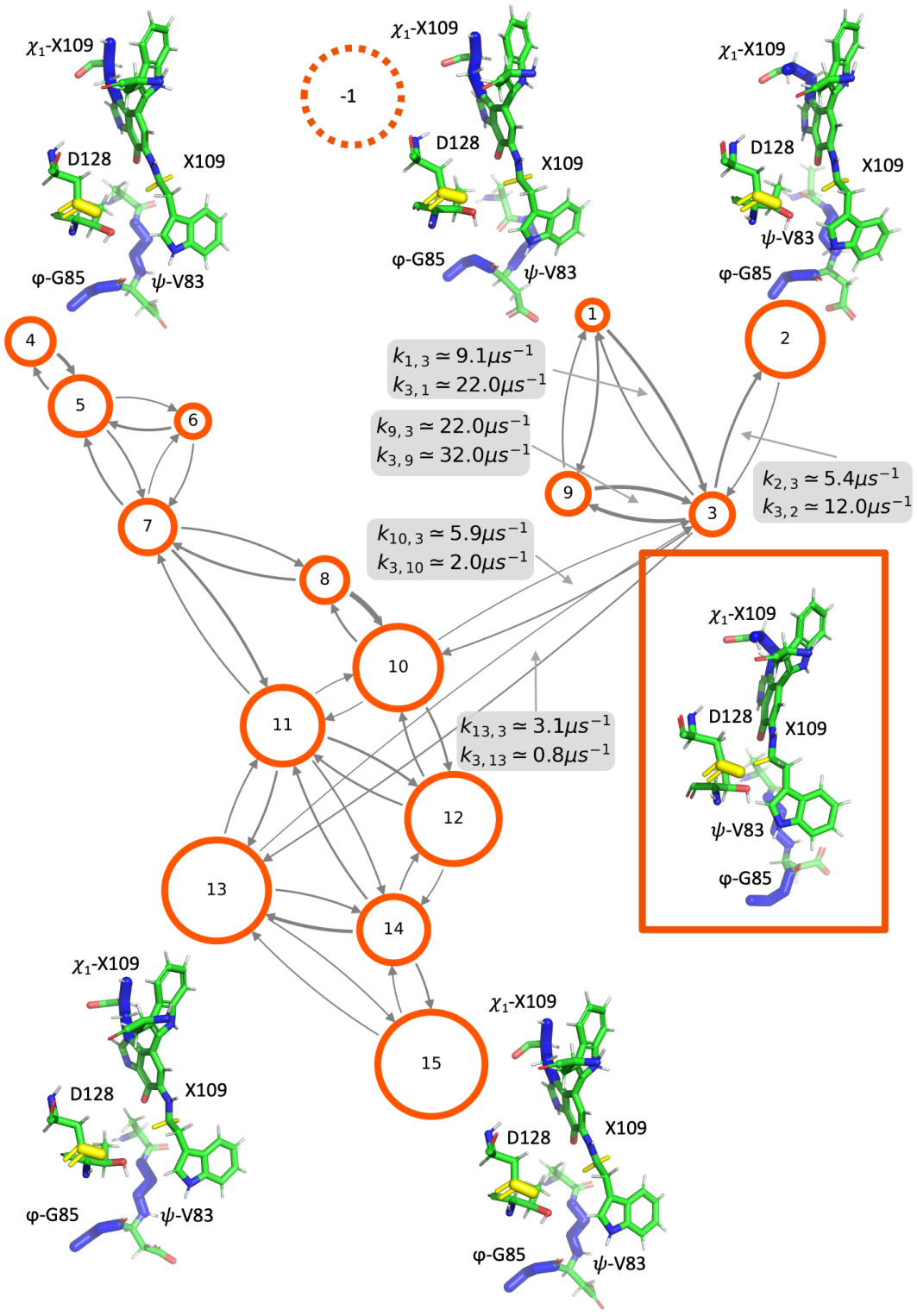


FIGURE 6.30: BASE CASE HMM AS A NETWORK. Each orange circle is a hidden state with area proportional to its stationary distribution. Arrows connect states i and i where the flux, $\tilde{\pi}_i \tilde{T}_{ij}$, is greater than 0.01%. The orientation of the states mirrors their position in the TICA plane. Representative structures are shown for $h_{-1}, h_2, h_3, h_4, h_{13}, h_{15}$. The most important TICA dihedrals are shown in blue. The yellow bonds are the C—OD1/OD2 on the Asp 128 residue (D128) and the CI—HI2/3 bonds on TTW 109 (X109). Rates of interconversion are shown for those states connected to h_3.

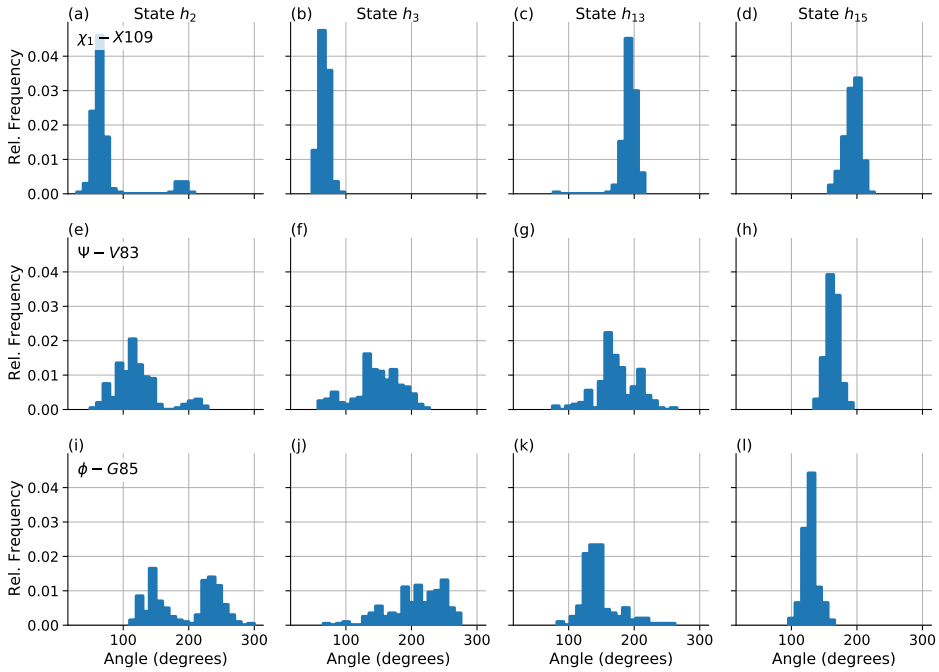


FIGURE 6.31: DISTRIBUTION OF IMPORTANT DIHEDRAL ANGLES IN THE BASE CASE HMM. The χ_1 dihedral angle of TTW 109 for hidden state 2, 3, 13, and 15 are shown in panels (a) - (d); the ψ backbone dihedral of Val83 in panels (e)-(h); and the ϕ backbone dihedral of Gly85 in panels (i) - (l). Angles are shown in degrees and have been shifted so that they lie in the range 0° to 360°.

(275 ns [53.4 ns–7400 ns]).

Sensitivity 2 state h_5 is structurally most similar to the base case state h_6. This was determined by looking at the distributions of RMSDs shown in figure 11.7. The RMSD distribution within each base case hidden state ('BC-BC') and between the base case and sensitivity 2 h_5 ('BC-S2') were compared to the distribution of RMSD within sensitivity 2 h_5 ('S2-S2'). For base case h_6 the three distributions were similar and the two states were determined to be the same. However, base case h_6 has a much shorter lifetime (30.4 ns [13.2 ns–110 ns]) and smaller stationary distribution (1.6% [0.1 ns–5.8 ns]). Despite these differences, the dominant timescale process in both base case and sensitivity two are similar (1210 ns [533 ns–6590 ns] for sensitivity 2 and 1180 ns [373 ns–8180 ns] for the base case). The difference in the VAMP-2 scores for the MSMs are small, but the coarse grained models of the base case and sensitivity 2 result in very different models. While this is expected of different features, the similar implied timescales and the similar VAMP-2 scores suggest that the being able to distinguish between these models will require further analysis and potentially more data collection.

6.6 Alternative analyses

An alternative to using the ICL to select the number of hidden states in a HMM model is to simply stipulate a two state model. This can be justified by the fact that this is the simplest model involving more than one conformational metastable states, and, that there is a separation in timescales between the dominant timescale and the remaining timescales shown in figure 6.21, panel (c). It is also the model stipulated in the extended transition state theory model of reference [220, 221], discussed in the introduction.

Some likely features of such a two state model can be inferred from the 15 state base case model shown in figure 6.27. First, the two new states (A and B) will be made up of unions of the existing 15 states. The sign of the dominant eigenvector (Ψ_2, panel (c)) gives an indication of how these states would be assigned to A and B, using a similar approach to PCCA+ for coarse-graining (see section 2.4.2). Ψ_2 is negative for state 2 (in orange, panel (a)) and positive for the remaining states. This indicates that the slowest dynamic process involves interconversion of state 2 and the remaining states. As a consequence, state A would consist of the state 2 and state B would consist of states 1 and 3 to 15. However, this results in a model which loses some important details, which the 15 state model brings out. Primarily, the DAD to OD2 (and to a lesser extent to OD1) for state 3 clearly shows that it is a distinct conformations which could be important for understanding reactivity (figure 6.29, panels (a) and (c)). For example, new molecular mechanics or QM/MM simulations could be seeded from this state to investigate other reactive paths and conformations. If this was subsumed into the larger state B, this detail would have been lost. Additionally, the TTW109 χ_1 dihedral angle is the dominant feature making up the first TICA component (i.e., its value changes left to right on figure 6.30). States 2 and 3 are clearly similar with respect to their values of this feature (figure 6.31 panels (a) and (b)) suggesting that they should should be grouped together, rather than separately in A and B. However, it is possible that the sign of Ψ_2 splitting states 2 and 3 is an artifact of the noise in the data, as evidenced from the large error bars in the implied timescales (see figure 6.27 panel (d) and figure 6.19). A two state model, while not immediately preferred, would be a clear cut candidate for further study.

The main drawback of the analysis thus far presented is that the final coarse-grained HMMs could not be validated. The reason for this is because of data removal as part of the Bayesian estimation algorithm used to estimate the HMMs. To summarise the method used to arrive at the base case Bayesian HMM discussed in section 6.4 (although this applies to all HMMs estimated): maximum likelihood HMMs with different numbers of hidden states were estimated and the ICL calculated for each one. The ICL selected an 18 state maximum likelihood HMM. The analysis method was then switched to Bayesian estimation, using PyEMMA, in order to accurately estimate the errors, or credible intervals, of the various HMM observables (e.g., implied timescales). The Bayesian estimation algorithm first removes frames from each trajectory which it estimates are highly correlated, in order to remove bias in the estimated HMM parameters. As a result, $95\,\%$ of the simulation data was removed. This data removal process is discussed in full in section 2.3.4.3

and is a standard technique in Bayesian analysis [252]. Because of this data removal, the new transition matrix in the Bayesian HMM had only 15 ergodic, or strongly connected, hidden states. The remaining three hidden states were effectively disconnected because transitions between the two groups of hidden states were observed only in the frames that were removed. This phenomena is discussed in section 2.3.4.2. The effect of data removal on the number of ergodic hidden states was dependent on the Markov lag time, meaning the definition of the HMM changed with τ. This meant the HMMs presented here could not be validated by the Chapman-Kolmogorov test.

An alternative to this process would have been to bootstrap maximum likelihood HMMs, thus mitigating the need to remove MD frames from the analysis. This would have kept the definition of the HMMs consistent at different values of τ and allowed these models to be validated using the Chapman-Kolmogorov test. To perform a bootstrap analysis, the discretized MD trajectories, used to estimate the MSM and HMMs, would be sampled with replacement N times, to create N different data sets. N would be chosen large enough that the resulting confidence intervals in model observables converged (N is typically $\sim$100–1000 [256]). For each of the N data sets a maximum likelihood HMM would be estimated and the implied timescales (and other observables) calculated. From the resulting distribution of implied timescales, 95 % confidence intervals could be calculated. The Chapman-Kolmogorov test would also be applied to these models to check the models for consistency with data. This avoids the need to remove frames from MD trajectories because the maximum likelihood estimate of the transition matrix elements is is not affected by correlations in the data.

6.7 Conclusions

AADH is an important system for studying the effects of enzyme conformational dynamics on catalysis due to the large and temperature independent kinetic isotope effect. Previous computational and experimental work determined the mechanism and estimated the free energy barriers using tryptamine as a substrate. Interpreted in the light of the full-tunneling models, the temperature dependence of the rate limiting step is due to the heavy atom reorganisation to achieve a tunneling ready state. The picosecond vibration that samples the donor acceptor distance in the tunneling ready state identified in reference [216] only introduces a negligible temperature dependence. The role of conformational dynamics in the full-tunneling models, the two-state transition state theory of [220] or other non-equilibrium proposals has still not been resolved.

The first aim of this work was to estimate the conformational landscape of AADH in its reactive (Schiff-base) state. The picture arising from the base case model shows the active site has a ensemble of short lived (20 ns to 300 ns) metastable states, with one state (state 3) with a significantly shorter donor-acceptor distance (approximately 4Å) than the other states. This state is a flux bottle neck state between two long lived states which interconvert on a timescale of approximately 1.2µs, primarily as the result of a rotation in the χ_1 dihedral angle on the TTW residue. The reason why

this process proceeds through bottle neck state has not been elucidated. Sensitivity 2 is a model with a similar, but smaller, value of VAMP-2 score for the corresponding MSM, but the picture arising does not back up that from the base case model. This either sheds doubt on the robustness of the conclusions from base case model, or on the ability of VAMP-2 score to differentiate between models with the amount of data collected here. The models presented here did not show any metastable states which could correspond to reactive or tunneling ready states. Therefore this work is of limited immediate use in resolving the role of conformational dynamics the reactivity of AADH.

The limitations of this work stem from issues in the simulation data. An error in the preparation of the simulation resulted in a missing an disulphide bond adjacent to the active site in the H chain. The conformations of the H and D chains were radically different, although surprisingly the H active site was more similar to the reactive conformations found in previous QM/MM studies. The effect of the large conformational change in the loop 92–108 in chain D and its effect on the dynamics of the active site was not investigated. If interactions from this loop with the active site change, the the assumption of stationarity when estimating the transition matrix is not valid. In addition, a number of unobserved residues were not modelled and there was moderate correlation between each trajectory's initial configuration, further limiting the generalizability of the simulations.

The second aim of this work was to critically assess the optimisation and model selection methods built up in chapters 4 and 5. The MSM building process was initialized by building an exploratory Markov state model and estimating a Markov lag time of $\tau = 2\,\mathrm{ns}$ and the number of dominant eigenvalues, $r = 4$. The response surface of an MSM of the active site, with four hyperparameters (χ, τ, m and n), was modelled with a Gaussian process. This highlighted some of the weaknesses of GPs for modelling these types of response surface. First, the conditional structure of the predictor space meant that the RMSD feature was not able to be incorporated into the response surface - a known problem for for GPs in this setting. Second, although standard model selection techniques were used to select the most appropriate GP kernel and input warping, the selected GP model fit the hyperparameter trial data set with varying success across each value of the protein feature. For three out of the four features the model fit the data well, however, the fit of the response surface for the best performing feature, the (ϕ, ψ, χ) dihedral angles, was less satisfactory. Third, the hyperparameters which failed to produce a converged MSM were ignored in the response surface modelling. This is ignoring valuable information and in future work this should be incorporated, although it is not clear how this may be achieved effectively without resorting to a different type of response surface model.

However, modelling the hyperparameter response as a GP had a number of practical benefits on top of those mentioned in chapter 4. First, the definition of the incumbent - best performing set of hyperparameters - uses information from all the other sampled trials through their influence on the mean and uncertainty of the GP. This is important as it decreases the influence of chance fluctuations in the measured response and increases the reproducibility of the result. Second, it allowed the relevance of hyperparameters to be rigorously defined and interpreted for both quasi-continuous

(i.e. τ, m and n) as well as categorical (i.e. χ) hyperparameters. This led to the conclusion that the the TICA lag time, τ, is more important in determining the VAMP-2 score than the number of microstates, n, in line with the findings from chapter 4. This will increase efficiency hyperparameter selection in follow up work on this system, and may generalize to other systems. Third, the relevance of the hyperparameters was able to guide efficient visualisation of the five-dimensional response surface, enabling suitable sensitivity analysis to be suggested. Making use of the relevance in this way will be useful for not only MSMs but for understanding and optimising other complex statistical and machine learning models.

Bayesian optimisation was used to optimise the MSM response, although this did not affect the final set of hyperparameters appreciably. Despite this, the optimisation step was important to confirm the convergence of the MSM response surface. By seeding the optimisation algorithm with fewer observations, Bayesian optimisation was able to find the maximum of the response surface with a smaller overall number of MSM evaluations. Although not conclusive, this work will hopefully encourage further research in this area to further reduce the amount of computational resources necessary for optimising MSMs.

The ICL was used to select the number of hidden states used to coarse grain the four MSM models (base case and three sensitivity tests), using a HMM approach. The selected HMMs were re-estimated using a Bayesian analysis in order to quantify the uncertainty in the model observables. This method is attractive because maximum likelihood models are relatively quick to estimate and the ICL requires negligible extra calculations. The more laborious Bayesian fitting procedure can then be reserved for models with higher probability of being useful. This meant that a large number of models could be quickly assessed for goodness-of-fit. However, this approach was not satisfactory for a number of reasons. First, the difference in the way the maximum likelihood and Bayesian HMMs are estimated means that the number of reversibly connected hidden states differed between the maximum likelihood and Bayesian approaches. This renders the interpretation of the ICL (as the model with the greatest evidence for observed and hidden state structure) less meaningful. Second, in the case of sensitivity models 1 and 3, the selected Bayesian models could not be estimated at all. Additionally, the base case and sensitivity 2 models could not be validated via a Chapman-Kolmogorov test. This is the result of the data removal process in the Bayesian estimation algorithm. An alternative analysis method using the Bootstrap procedure was suggested which mitigates this data removal process. Lastly, an two state model was considered and its properties inferred from the base case model already presented. While a two state model would be justified, it may be too coarse-grained to capture some of the potentially important features of the 15 state base case model.

A program of future work to address the limitations of both the simulation data and MSM construction issues (beyond that already addressed in section 6.6), as well as to extend this work is presented in chapter 7.

CHAPTER 7

CONCLUSIONS AND OUTLOOK

7.1 Summary of main findings and improvements

Chapter 3 demonstrated the use of Markov models to explain deviations to the Stoke-Einstein (S-E) equation of water diffusion in a sucrose matrix. Compared to the workflow presented in later chapters, a simplified modelling approach was taken. After inspection of the simulation data, the natural MSM basis set used to model the water dynamics was microstates clustered in the space of the Cartesian coordinates of the water molecule. Heuristics from the from the literature guided the other modelling choices and no variational optimisation of basis sets, of the kind described in chapter 4, was performed. This meant that the modelling could be done at scale - 8000 different MSMs were estimated and used to partition, classify and analyse the MD data. The results showed that $11.8\,\%$ of the time the water was trapped in cavities defined by the sucrose matrix undergoing reversible hopping dynamics with an average barrier height of $6.42 k_B T$.

To develop this work further a number of improvements can be made. First, the proportion of time spent in local equilibria could be refined by more accurately partitioning each MD trajectory into reversible/non-reversible sections. Instead of a single sucrose position decorrelation time (which determined the $1\,\mathrm{ns}$ time-slices used in chapter 3), a local estimate can be used by averaging the autocorrelation function over small range of values $t - \Delta t < t < t + \Delta t$:

$$\mathrm{R}_{X,X}(t') = \mathbb{E}_{t \in [t \pm \Delta t]} \left[X_t \bar{X}_{t+t'} \right] \tag{7.1}$$

and choosing the longest time such that $\mathrm{R}_{X,X}(t') > 0.8$ (or some other threshold). By repeating this process along the whole trajectory an 'adaptive' partition of trajectories into stationary would be achieved. Second, using this more accurate partition, the irreversible timescales could be estimated using Koopman models [96]. With both sets of timescales estimated a more mathematically precise comparison of the S-E and cavity dynamics model can be made. Third, using more complex basis

sets involving water-sucrose interactions could be variationally optimised using the methods of chapter 4, potentially shedding more detailed light on the microscopic mechanism of transport between the cavities.

Chapter 4 demonstrated how to use Bayesian optimisation of model hyperparameters, a technique common in machine learning, for an MSM of alanine dipeptide. Specifically, Gaussian processes regression was used to model the response surface of an MSM with respect to its hyperparameters and Bayesian optimisation used to optimise the number of microstates, n and the continuous feature, χ. Gaussian processes regression modelled the response surface well. Calculation of the hyperparameter relevance, quantified what was already apparent from visualisation of the response surface, namely: that the VAMP-2 score was insensitive to the number of microstates, and, for different protein features, the shape of the response surface was similar. An interpretation of the hyperparameter relevance for categorical variables which extended the usual definition [306]) was presented. Both the (ϕ,ψ) dihedral angles *and* the heavy atom positions captured the slow relaxation processes equally well. Bayesian optimisation was then used to optimise the hyperparameters and it was shown that 50 seed observations was needed to initialise the Bayesian optimisation algorithm. However, this did not improve the VAMP-2 score, but did provide a useful convergence check on the hyperparameters already selected.

The main limitation of this study is that the response surface was too simple and did provide adequate complexity to test the Bayesian optimisation algorithm (proven in other machine learning model studies [135, 306]). In addition, the irrelevance of the number of microstates, n, was not in line with other MM studies [94, 97, 194]. This counter-intuitive result needs to be verified. To do this, the search space of n should be shifted from $[10,1000]$ to $[2,100]$; MD data with a higher temporal resolution data used; and the search space expanded to include different clustering algorithms. In addition, this should be replicated on different benchmark systems, e.g., the Villin headpiece and β-lactamase used in reference [168] or the twelve fast folding proteins [93] used in other studies [92, 246, 253]. Using these extended benchmark systems, the demonstration of Bayesian optimisation (BO) and the test of the number of required seed observations could be improved by: i) increasing the number of Bayesian optimisation steps from 10 to 100, or until convergence in the response surface, and ii) repeating the optimisation a larger number of times to gain a statistical insight into its efficacy, and a more generalizable estimate of the number of seed observations. Working with a set of well-converged simulations of larger proteins, would increase the generalizability of the conclusions about hyperparameter relevance, and provide a more robust test of Bayesian optimisation than just alanine dipeptide. To extend this work even futher and bring statistical modelling of biomolecular systems ever closer to machine learning best practice, the use of response surface models which take into account conditional hyperparameter search spaces such as tree Parzen estimators or Random forests could by tested.

Chapter 5 drew on ideas from the mixture model community to demonstrate the use of the integrated classification likelihood (ICL) and other information criteria in determining the appropriate

number of hidden states in a HMM. The benefits of these type of model selection metrics are that they are easy to calculate and do not rely on being able to resolve clear gaps in the eigenvalue spectrum. The ICL was able to correctly determine the number of metastable states in an HMM of the four well Prinz potential for 3 out of 5 lag times tested, while selecting only one extra state for the remaining 2 lag times. The remaining criteria overestimated the number hidden states by a considerable margin. The main limitation of this work is the limited amount of benchmarking. A number of further steps are needed to gain a better estimate of the generalizability of the ICL. First, bootstrap the simulation data and calculate the selected number of hidden states on each bootstrap sample. This would give probabilities of the each criteria selecting the correct number of states, rather than the current single point estimates. Second, and still using the Prinz potential simulation data, calculate performance of each criteria with differing amounts of data. This would allow the performance of each criteria under the more realistic condition of limited data to be assessed. Third, more benchmark systems are needed, both 'toy' models with well defined numbers of metastable states e.g. the 2D triple well from reference [169], the model 2D potential in reference [188] used to benchmark Bayes factors for MSMs, and Müller potential in the AIC and BIC benchmarking in reference [194]. Other, more realistic, benchmark systems should also be used, such alanine dipeptide, the Villin headpiece and β-lactamase (used in reference [168]) and the Fip35WW (used in reference [194]). Lastly, the results of the ICL could be compared the full Bayes factor method from reference [188].

Chapter 6 used the response surface and Bayesian optimisation methods from chapter 4 and the model selection techniques of chapter 5 to create a HMM of the active site of AADH. A simulation data set of AADH was created and the response surface using an expanded set of hyperparameters was estimated and optimised. A number of sensitivity models were suggested from inspection of the optimised response surface. HMMs were estimated using the number of hidden states suggested by the ICL. The most relevant parameters for determining the VAMP-2 score were the TICA lag time and the least relevant the number of microstates. The optimal continuous feature was the backbone and residue dihedral angles. Coarse graining the optimal MSM using a Bayesian HMM and choosing the number of hidden states with the ICL revealed a complex network of states. The slowest relaxation process of approximately $1\,\mu s$ involved transport between two unreactive states through a flux bottleneck state. This bottle neck state had the shortest average donor-acceptor distance and was thus deemed the most 'reactive'. The second sensitivity test (and the only sensitivity test to converge a HMM) did not confirm these results and showed a qualitatively different story albeit with similar relaxation timescales. Due to the effective lack of data and the Bayesian estimation algorithm, neither of the HMMs could be validated. This work failed to find evidence for the hypothesis from reference [221] that there are two distinct, reactive conformational states.

The main drawback of this work is the AADH data set which must be corrected and expanded before any inferences can be drawn. First, the missing disulphide bridges must modelled correctly and the simulation cell re-equilibrated. Second, the sampling trajectories must be initialized from

independent starting structures by minimizing and re-equilibrating structures taken from a seeding trajectory. Third, the source of the large conformational changes both in the loop structure adjacent to the active site and the tail residues must be investigated. This will ensure that the sampling is taking place in conformations relevant for the rate determining reaction. The amount of simulation data should be determined by monitoring the convergence of implied timescale for optimized MSMs (using the methods of chapter 4), or through an adaptive sampling method [403].

7.2 Lessons learned

The objectives of this thesis were to investigate whether model selection and optimisation methods, common in the machine learning and statistic community, could be used to create robust MSMs in a more transparent and efficient way. Two systems were studied utilising two different workflows. The first was water diffusion in secondary organic aerosol particles which were modelled using using a simplified Markov model construction workflow. MSM and HMM parameters (number of basis states, hidden states, essential degrees of freedom) were selected using heuristics from existing literature and visualisation of the simulation data. The second was the conformational dynamics of AADH, an important enzyme in the debate surrounding tunneling and the role of dynamics in enzyme catalysis. The workflow used here was more complex and consisted of modelling and optimising the response surface of an Markov state model, creating sensitivity tests, and using statistical model selection to select a final coarse-grained description with a hidden Markov model. A number of lessons and themes can be drawn from comparing these two approaches.

Lesson one. The more complex Markov modelling workflow, involving modelling and optimising response surfaces, is not necessary to produce scientific insight. The results of the simplified Markov model workflow produced interpretable and novel insights into the mechanism of water diffusion in organic aerosol particles. However, this does not rule out the possibility that a more complex optimisation procedure would bring benefits. To test this, a range of different features describing the water-sucrose system could be optimised using the methods in chapters 4 and 5.

Lesson two. The choice of feature is important but the VAMP-2 scores of MSMs are largely insensitive to other hyperparameters. Visual inspection of the response surface of alanine dipeptide, and the calculation of the hyperparameter relevance for AADH, showed that the VAMP-2 score was insensitive to the values of the number of microstates and TICA parameters. The choice of feature was important, but in both alanine dipeptide and AADH, more than one feature was identified as being optimal.

Lesson three. Bayesian optimisation is not necessary for optimising low relevance hyperparameters. Bayesian optimisation was applied to the response surface of both alanine dipeptide and AADH and failed to improve the VAMP-2 scores in both cases. However, when seeded with an appropriate amount of data, the Bayesian optimisation algorithm provided a convergence check on the optimum value of the VAMP-2 score.

Lesson four. Modelling and visualising the response surface of a model allows the creation of principled sensitivity tests. In order to test the conclusions of a statistical analysis, results should be compared after changing key modelling choices. Sensitivity tests 2 and 3 for AADH were devised after inspection of response surface showed how different features and TICA lag times affected the VAMP-2 scores.

Lesson five. MSMs with similar VAMP-2 scores produce different results. Three models with different hyperparameters were produced based on their similar VAMP-2 scores for AADH[1]. However, the results from two of them did not produce similar descriptions of conformational dynamics. However, this could be the result of inadequate simulation data.

Lesson six. The integrated complete data likelihood criterion is a promising model selection criterion for selecting the number of metastable macrostates. It selected the appropriate number of metastable macrostates in the model system used, but the test was not thorough and it should be tested further. However, it is simple to calculate, and its interpretation is in line with the goals of a hidden Markov model analysis, namely that coarse-graining is a trade-off between crisp partitioning of microstates into macrostates and modelling the dynamics accurately. However, it should be noted that this only applies to data in which the dynamics are metastable. When it is important to model transition regions, the HMM approach to coarse-graining is not appropriate.

7.3 Outlook and further work

The outlook for further work builds on these lessons and the specific findings of chapters 3 and 6 in three areas.

First, Markov models can and should be used for areas outside of large biomolecule simulations. The mechanism of water diffusion suggested by this work is not proven generally and needs investigation in both the other saccharide systems experimentally studied here, as well as through other amorphous materials such as those found in pharmaceutical delivery systems [404].

Second, the classification likelihood used in the ICL selection metric, could be used to extend the work of reference [188] to create a *classification* Bayes factor. This would provide an alternative to the current Bayes factor for judging the quality of a coarse-graining method and number of metastable states. The information gained from this type of Bayes factor would be to more accurately (than the approximate ICL) determine how well a given coarse-graining crisply partitions microstates into macrostates, while both preserving Markovian dynamics and taking into account over-fitting. Like the full Bayes factor method of reference [188], however, this would require much more computational effort than the ICL. However, Bayesian variational inference methods [405], which replace the expensive sampling of traditional Bayesian estimation with approximations of the Bayesian posterior distribution, could be used as an approximate method for calculating both types of Bayes factor.

[1] Sensitivity test 1 changed the Markov lag-time which is not a hyperparameter.

Third, the metrics used to score Markov state models need further investigation. Lesson five suggests that the VAMP-2 score may hide important differences between MSMs built with different hyperparameters. If this is accurate and holds generally, this would be an example of the *Rashomon* effect: where models which perform similarly according to some metric (e.g., predictive performance, VAMP-2 score), differ in interpretation [318]. If MSMs do exhibit the Rashomon effect then this would have important consequences for the interpretation of simulation data. For example, if two models with similar VAMP-2 scores describe different dynamical processes then this needs to be taken into account when reporting results. This would also call for further tests (experimental or computational) to determine which picture is correct. A test for the Rashomon effect would be to estimate models using different protein features and compare the results, using well converged simulation data, e.g., the benchmark twelve fast-folding proteins [93].

Fourth, response surface methods and Bayesian optimisation can be used to facilitate a robust Markov modelling workflow. The users of Markov models for understanding biomolecular systems are not necessarily experts in statistical model development. This has motivated the publication of user friendly packages for creating arbitrary Markov models [253, 406]. These packages facilitate a workflow where only a handful of different sets of hyperparameters are used, such as the simplified workflow from chapter 3. The next stage of Markov model development could be to create not a handful of MSMs, but the entire MSM response surface for a given set of simulation data. This would allow the creation of sensitivity tests for testing the robustness of statistical inferences. While the lessons learned from this thesis suggest that only the protein feature is important for determining the VAMP-2 score, the response surface methodology could incorporate other metrics, which may be more sensitive to other hyperparameters. If the response surface is sensitive to the hyperparameters, Bayesian optimisation could be used to optimise this surface. Steps towards this have been taken with the creation of Osprey [148]. In addition, the author of this thesis is currently engaged in a code re-factoring of this package to accommodate the work of chapter 4. In future this may allow automatic optimisation of hyperparameters as well as understanding counter-intuitive features via their response surface. For example, solvent degrees of freedom are important but often ignored [407]. The distance metric for incorporating solvent degrees of freedom in reference [407] does not have the salience of, say, a dihedral angle. However estimating and optimising the response surface with respect to its hyperparameters (the number of number of solvent molecules incorporated and the width of its distance kernel, σ) can give an intuitive understanding of its effect on explaining the kinetic variance. Beyond the realm of Markov models, understanding machine learning models through their response surface could help in breaking down technical barrier between non-expert users and the ever increasing set of methods for understanding molecular simulations [408].

CHAPTER 8

CAVITY DYNAMICS OF WATER IN A SUCROSE MATRIX

Appendix 8 reproduces the supplementary information of reference [109]. See chapter 3 for a statement on the contributions to this work.

8.1 The experimental procedure for determining timescales of water transport

In order to determine evaporation and condensation timescales for binary systems that are mixtures of water and organic compounds, they were subjected to a single transition step of relative humidity (RH) across different particle size and between different pair of target RH.

An individual binary solution aerosol droplet ($3\,\mu m$–$10\,\mu m$ radius) is captured at high RH/dilute saccharide concentration in a gradient-force optical trap (optical tweezers) formed by a tightly focused laser beam (wavelength $532\,nm$) [285, 409]. After a period of conditioning at an elevated RH (typically $70\,\%$–$80\,\%$), the mixing ratios of humidified and dry nitrogen are altered to drive a change in RH in the trapping cell. Measurements proceed either through small downward (or upward) steps in RH of $10\,\%$ or following large changes of $30\,\%$ or more. The moisture content of the droplet responds accordingly, attempting to remain at a water activity equal to the surrounding RH, driving a change in particle size through the evaporation or condensation of water.

The time-dependence in the particle size and refractive index (hence composition and moisture content) are inferred (with sub-nm precision for size) from the shifts in the wavelengths of whispering gallery modes apparent on the Raman O-H stretching band of water [410–412].

Figure 8.1 shows one example of KWW fitting for the aqueous-sucrose system. The response in particle size is inferred from the wavelength shift indicated by the red line, starting at a particle size of $3650\,nm$ and a wavelength of $655.4\,nm$ following an initial period of $3\,h$ conditioning at $30\,\%$ RH.

From this point, the KWW equation was applied to fit the evaporation step (RH 30 % to 5 %) as shown by the blue line. The fitting of the condensation step (RH 5 % to 30 %) is shown by the green line. The timescales of water transport for each process were 1841 s and 2566 s, respectively. As the fitting results demonstrate, the KWW equation is a suitable method to determine the relaxation process.

It has previously been shown that stretched exponential relaxation reflects the departure from a single-exponential relaxation in size as the particle becomes increasingly viscous [287],

$$F(t) = \frac{r(t) - r(t=\infty)}{r(t=0) - r(t=\infty)} = \exp\left[-\left(\frac{t}{\tau}\right)^{\beta}\right] \tag{8.1}$$

where t is the experimental time, τ is the timescale of relaxation and β characterizes the shape of the multi-exponential time-dependence. The response function, $F(t)$, characterizes the fractional relaxation from the initial particle radius r at $t = 0\,\mathrm{s}$ through to the final equilibrated radius at $t = \infty$ following the change in RH.

This KWW fitting process has been carried out using customized LabVIEW software.

8.2 The effect of particle size on equilibration time for six binary organic systems

Before moving on to fit the compositional dependence of diffusion constants, we consider first the qualitative dependence on droplet size. The experiment conditions are divided into two main categories for evaporation and condensation in this study: a low viscosity transition (RH 50 % to 30 % and RH 30 % to 50 %) and a high viscosity transition (RH 30 % to 5 % and RH 5 % to 30 %). The six panels in figure 8.2 show measurements for six binary systems of aqueous-glucose, aqueous-sucrose, aqueous-trehalose, aqueous-raffinose, aqueous-maltose, and aqueous-levoglucosan. All measurements for the six binary organic mixtures clearly show different trends with respect to initial or final RH, but the characteristic timescale of water transport (τ) clearly also shows a particle size dependence during evaporation or condensation.

In the high relative humidity region, transitions RH 50 % - 30 % - 50 % (i.e., the evaporation step of RH 50 % to 30 % and condensation step of RH 30 % to 50 %), the droplets of maltose showed the largest value of timescale of water transport (τ) over the other organic particles, see figure 8.3. Moreover, the characteristic timescale increases with increasing particle size for every binary organic system. For example, the characteristic timescale of raffinose is 188 s for 3332 nm, 265 s for 4110 nm, and 313 s for 4924 ns. Comparing similarly sized particles of $\sim 4\,\mu\mathrm{m}$ during an evaporation step, the organic particles show $514\,\mathrm{s} \pm 31\,\mathrm{s}$ for maltose (4114 nm), $250\,\mathrm{s} \pm 23\,\mathrm{s}$ for raffinose (3967 nm), $157\,\mathrm{s} \pm 13\,\mathrm{ns}$ for trehalose (4101 nm), $59\,\mathrm{ns} \pm 19\,\mathrm{s}$ for sucrose (4016 nm), $29\,\mathrm{s} \pm 3\,\mathrm{s}$ for glucose (3871 nm), and $32\,\mathrm{s} \pm 4\,\mathrm{s}$ for levoglucosan (3746 nm). The ordering of the timescales for all particles sizes in the range of 3 µm–6 µm show the same tendency: maltose > raffinose > trehalose > sucrose > glucose ≥ levoglucosan, as shown in figure 8.3. In this region, raffinose and

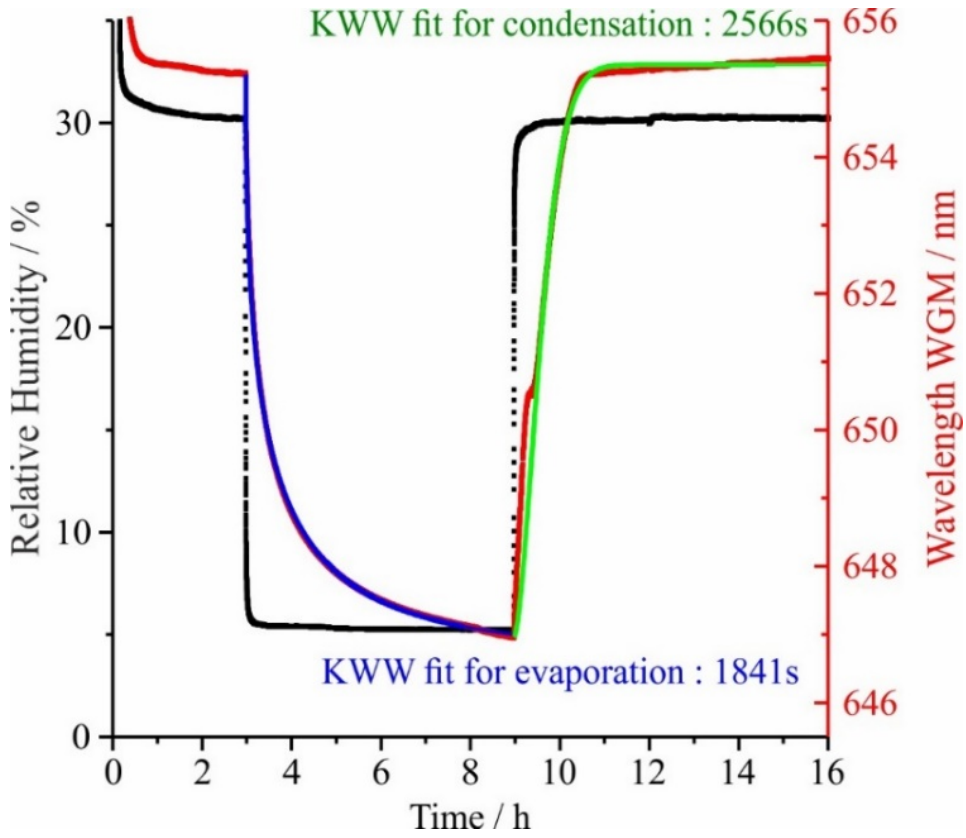


FIGURE 8.1: KWW FITTING FOR THE AQUEOUS-SUCROSE SYSTEM. On the binary-sucrose system, RH transition is from 30 % to 5 % for the evaporation step, and 5 % to 30 % for the condensation step. The black line is RH profile and the red line is one of WGMs change. The blue and green lines are fitted by KWW equation (equation 8.1). The timescales of evaporation and condensation are 1841 s and 2566 s, respectively.

maltose particles have experienced the glass transition at RH ~ 53 % and RH ~ 32 % in the ambient temperature [285, 288]. However, trehalose, sucrose, glucose and levoglucosan do not pass through the glass transition RH until a much lower value. The molecular diffusivity of water in each organic system will be treated in the next section.

For transitions at low relative humidity, notably for RH 30 % to 5 % and 5 % to 30 %, which is the high viscosity region for all compounds, the characteristic timescales of water transport of maltose particles have the longest water transport timescales in the low humidity region. Comparing similar sized particles of ~ 4 µm during an evaporation step, the organic particles showed the timescales as follows: 4062 s ± 55 s for maltose (3745 nm), 2035 s ± 23 s for sucrose (3809 nm), 1447 s ± 19 s for raffinose (3952 nm), 1342 s ± 17 s for trehalose (4040 nm), and 1009 s ± 10 second for glucose (3871 nm). Levoglucosan cannot be measured in this RH region with particles crystallising at higher RH. The ordering of the timescales for all particles sized in the range of 3 µm–6 µm show the same tendency; maltose > sucrose > raffinose > trehalose > glucose, as shown in figure 8.2. In this low RH region, raffinose (RH ~ 53 %) [285, 288], sucrose (RH ~ 23 %) [285, 288], trehalose (RH

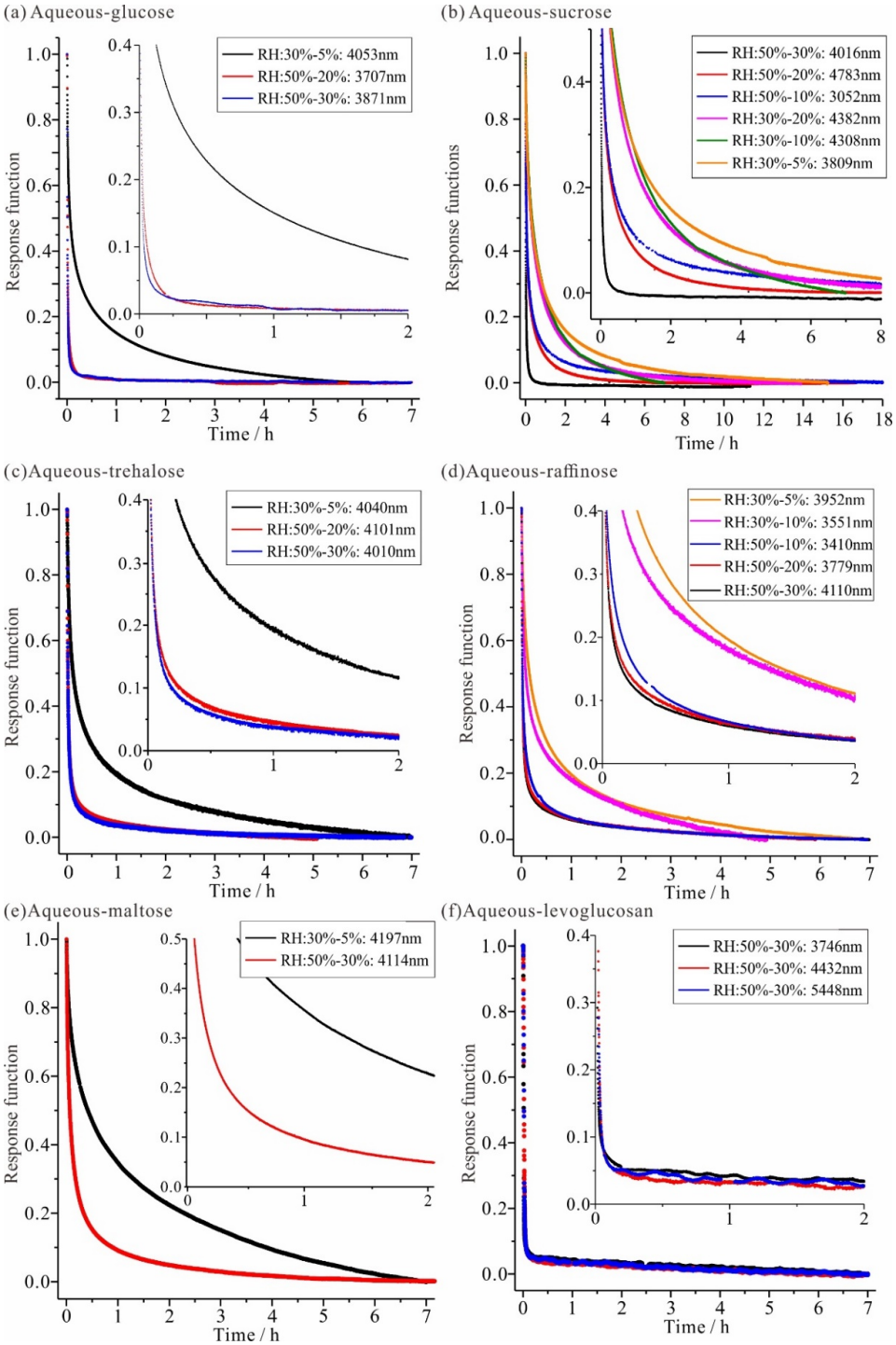


FIGURE 8.2: BINARY MIXTURE RESPONSE FUNCTIONS. Examples of the response functions for size changes of six binary mixtures particles following a step change in RH

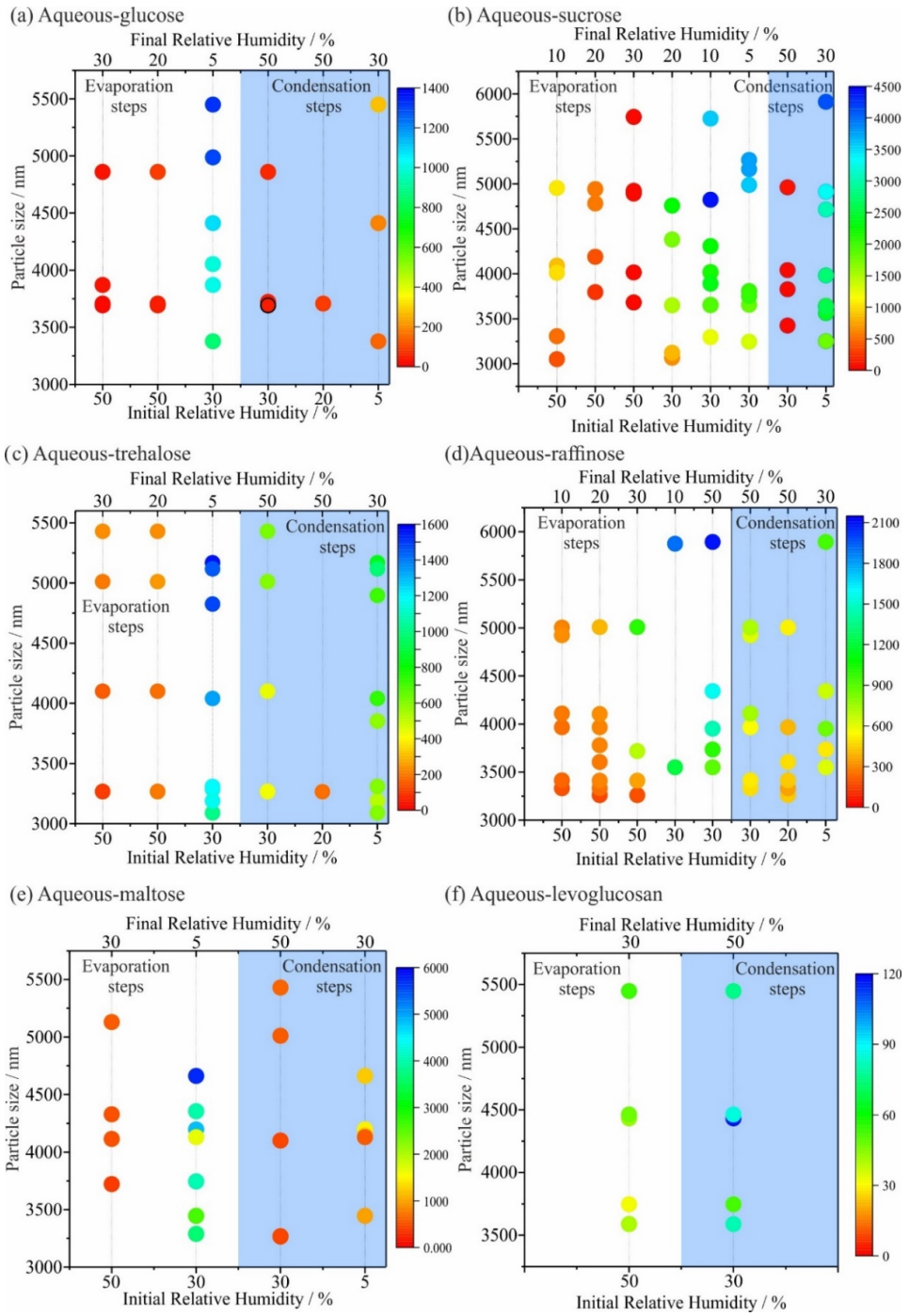


FIGURE 8.3: TIMESCALE OF WATER TRANSPORT (τ) FOR EVAPORATION STEP AND CONDENSATION STEPS. The values are also reported numerically in table 8.1. The evaporation and condensation timescales are determined by KWW function, and particle size is calculated from the droplet Raman signal by the proprietary LARA software. Compositions are: (a) binary aqueous-glucose (b) binary aqueous-sucrose (c) binary aqueous-trehalose (d) binary aqueous-raffinose (e) binary aqueous-maltose (f) binary aqueous-levoglucosan

$\sim 22\%$) [285, 413], and maltose (RH $\sim 32\%$) [285, 414] pass through the glass transition RH at ambient temperature except glucose, and glucose shows the smallest timescale for water transport. The characteristic timescale of water transport can be explained by water diffusion in the organic particles. All water transport experimental data go into Fi-PaD diffusion simulation.

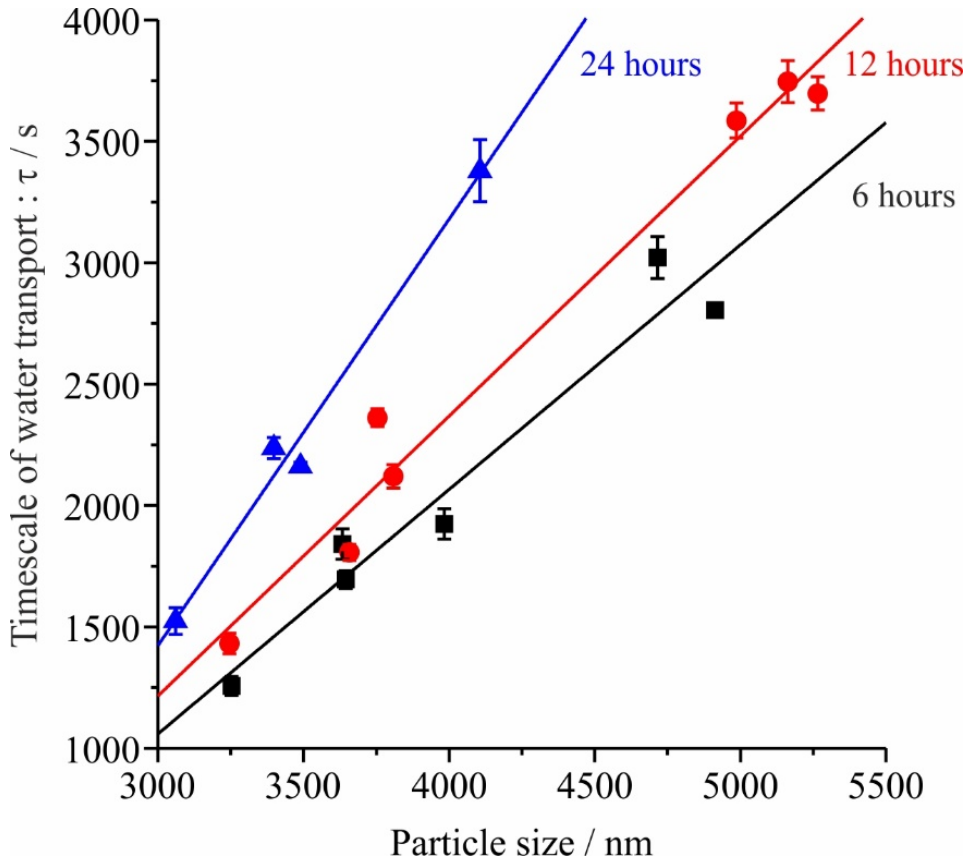


FIGURE 8.4: TIMESCALES OF WATER TRANSPORT. Timescale of water transport (τ) for each condensation step of sucrose, RH change from 30 % to 5 % then back to 30 % after drying. Particles of black squares experienced 6 hours drying time at RH 5 %, particles of red circles experienced 12 hours drying time at RH 5 %, and particles of blue triangles experienced 24 hours drying time at RH 5 %. Error bars were calculated a variation of $\beta \pm 0.1$.

8.3 Equilibration time variation with "wait time" Effect

The relaxation dynamics for condensation processes are dependent on the "wait time" (i.e., also referred to the aging of the particle). Figure 8.4 shows the impact of "wait time" on the relaxation timescale of water transport in a binary water-sucrose system. In this case the relaxation timescale is the timescale for the re-condensation of water following a period of drying of varying time (the "wait time"). Three different experimental conditions were studied. Sucrose particles that are indicated in black squares were dried for 5 hours at RH 30 %, and then the particles were held at RH 5 % for 6 hours. For comparison with this data set, sucrose droplets, indicated by red circles and blue triangles, were held at 12 and 24 hours at RH 5 %. After 6, 12, and 24 hours at 5 % RH, the RH is restored to the initial level 30 % RH. Several sucrose particles across a range of particle size ($3\,\mu m$–$5\,\mu m$) were studied at each RH transition. Figure 8.4 shows that when the particles experienced a long wait time, the relaxation timescale of water transport (τ) increased during re-condensation following

the increase in RH to the initial level. Depending on wait time under dry conditions, the water content varies both in magnitude and spatially before the RH is increased [287]. When increasing RH again, water vapor immediately condenses onto the particle surface, and leads to a unique level of heterogeneity which varies with the timescale over which moisture has been removed and the particle dried [287, 415]. The KWW equation was used to determine the timescales for the re-condensation step and the dependence on particle size is shown in the figure. When the particle size is larger, the timescale for re-condensation is also greater. The three trend lines in the figure, fitted by a linear equation, clearly indicate the relation between drying time, particle size and re-condensation timescale. The significance of the wait time is that when particle returned to the initial state from dry condition, the condensation time to return the particle to equilibrium with the surroundings increases as the wait time increases.

8.4 Fickian diffusion modelling (Fi-PaD model) for determining diffusivity of water in aerosol particles

In order to determine diffusivity of water in organic mixtures, a recently developed Fi-PaD model is applied to experimental data in this work. This was achieved using the same WGMs as were selected to fit KWW functions to.

O'Meara et al. developed a partial differential model, called Fi-PaD model, comparing it with two other diffusion models referred to as the ETH model and KM-GAP model (kinetic multi-layer model of gas-particle interactions in aerosol and clouds) [416]. The Fi-PaD model is established using Fick's second law. The Fi-PaD model uses three initial assumptions below:

1. An aerosol particle is spherical.
2. A spherical particle is divided into inner concentric shells.
3. A surface shell immediately reaches an equilibrium state with gas phase RH.

According to these assumptions, an aerosol particle consists of a number of concentric shells within the particle bulk in the Fi-PaD model. In this thesis, the number of shells is 400 and the resolution of a shell is around $10\,\mathrm{nm}$ for $4\,\mu\mathrm{m}$ radii of particle.

The Fi-PaD model is used to provide a forward simulation of the time-dependent size and response function following the step change in the gas phase RH change [111, 416]. A first guess for the water activity dependence of the diffusion coefficient is assumed. The diffusion coefficient of water in the mixture, D_w, is assumed to follow a Vignes form [111, 416]:

$$D_{\mathrm{w}}(x,\alpha) = D_{\mathrm{w,w}}^{x\alpha} \times D_{\mathrm{w,org}}^{(1-x\alpha)} \tag{8.2}$$

where x is the mole fraction of water, $D_{\mathrm{w,w}}$ is the known and limiting value of the diffusion coefficient of water in pure water, $D_{\mathrm{w,org}}$ is diffusion coefficient of water in pure solute at infinite dilution of

water. α is often viewed as analogous to an activity coefficient, given by:

$$\ln \alpha = A(1-x)^3, \tag{8.3}$$

where A is a temperature dependent parameter. Using Fick's second law, the model simulates the concentration change of every shell in the particle by calculating the approximate diffusional mass flux.

All measurements are performed at room temperature ($20^{\circ}\mathrm{C}$). Thus, two fit parameters (A and $D_{\mathrm{w,org}}$) are varied independently to achieve the best fit to the time-dependent size. For the example shown in figure 3.1 panel (a), the best-fit water activity dependence of the diffusion coefficient of water in the mixture is shown in figure 3.1 panel (b).

8.5 Holographic optical tweezers

Holographic optical tweezers (HOT) are used to initiate the coalescence process between two aerosol particles and to, thereby, infer the particle viscosity at a particular gas phase RH/particle moisture content [285]. In order to catch multiple droplets, the optical additional components are required and are shown in figure 8.5. The most significant is the inclusion of a Spatial Light Modulator (Holoeye SLM LC-R 2500, twisted nematic, liquid crystal on silicon) on the right side and a high bit-rate oscilloscope (LeCroy Wavesurfer 454) with photo detector (Thorlabs DET 110) on the left side, as shown in figure 8.5. The trapping beam is a continuous wave Nd:YVO4 laser at $532\,\mathrm{nm}$ (maximum output is $3\,\mathrm{W}$, Opus, Laser Quantum). After the laser output, the trapping beam is a vertically polarised by a half-wave plate, and it directly passes through a telescope which adjusts beam size. The SLM is used to split the beam. A beam expansion telescope and inverted microscope objective generate the optical traps to capture aerosol particles. Using the half-wave plate, the power division between the SLM and a beam dump can be controlled. In the trapping cell, multiple droplets can be held by the split beam. During experimental measurements, particle images are recorded by camera, and Raman spectra are recorded by the spectrograph. Light scattering patterns following the initiation of coalescence event are recorded by the photo detector.

8.6 Viscosity of saccharide solutions aerosol particles

Previously,the procedure for recording the viscosity of aerosols containing saccharides, alcohol, di- and tri-carboxylic acids has been described, reporting viscosity measurements as a function of RH [120, 285, 409]. Briefly, pairs of aerosol particles of identical composition are captured in parallel optical traps formed using an holographic optical tweezers arrangement. The particles are conditioned at a fixed RH (in the range $<5\,\%$ to $>90\,\%$ RH) for a period of time that can extend to hours. It has been shown that this allows sufficient time for the particles to achieve a moisture content in equilibrium with the gas phase RH and a uniform homogeneous composition/viscosity [285].

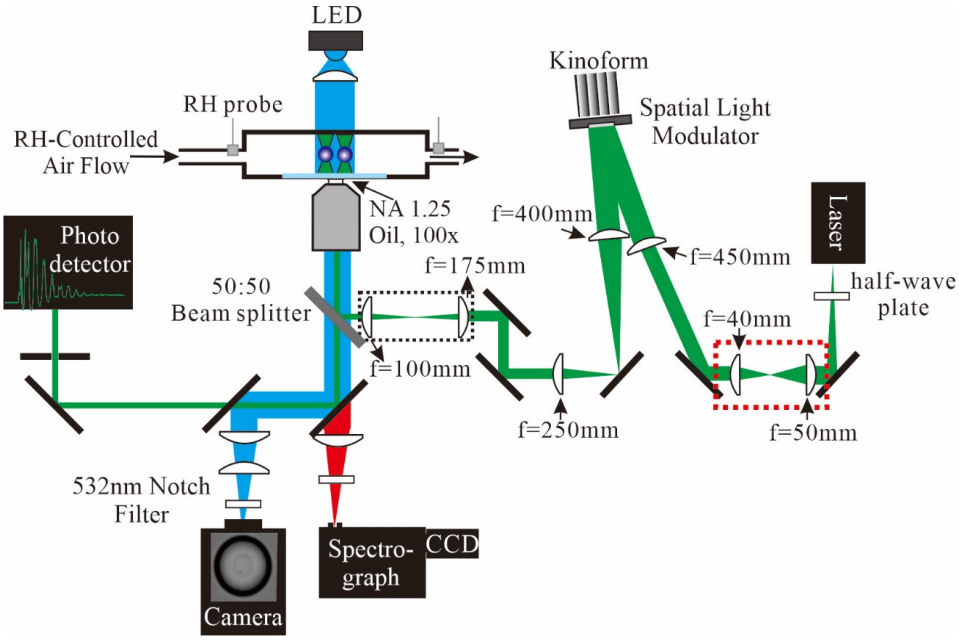


FIGURE 8.5: SCHEMATIC REPRESENTATION OF HOLOGRAPHIC OPTICAL TWEEZERS (HOT). The camera provides particle images, Raman spectrum is recorded by the spectrograph, and light scattering pattern is recorded by the photo detector [285].

Coalescence is initiated by beam-steering, merging the pair of optical traps. At low viscosity ($< 10\,\mathrm{Pas}$), the particles may merge and relax in shape on a timescale less than $1\,\mathrm{ms}$, a process that is monitored by measuring the relaxation in the back-scattered light from the optical trap. At higher viscosities, direct brightfield images can be recorded and the distortion in shape followed by determining the aspect ratio of the composite particle. The time-dependent decay in the aspect ratio can extend from milliseconds to $\gg 10000\,\mathrm{s}$, extending the viscosity range that can be measured from $1\,\mathrm{Pas}$ up to $10^8\,\mathrm{Pas}$–$10^9\,\mathrm{Pas}$ [285]. The viscosity is then inferred from the relaxation timescale, τ, assuming over-damped creeping fluid flow, the droplet radius and an estimate of the surface tension, σ:

$$\tau = \frac{\eta r}{\sigma} \tag{8.4}$$

A comparison of the viscosity trends for saccharides from mono- to tri- structures is provided in figure 8.6. There is a clear systematic trend toward higher viscosity as the composition progresses from a monosaccharide to a trisaccharide, depending on chemical structure, and molecular weight, with an order: glucose ($180.16\,\mathrm{g\,mol^{-1}}$) < sucrose ($342.30\,\mathrm{g\,mol^{-1}}$) < trehalose ($342.296\,\mathrm{g\,mol^{-1}}$) < maltose ($342.30\,\mathrm{g\,mol^{-1}}$) < raffinose ($504.42\,\mathrm{g\,mol^{-1}}$).

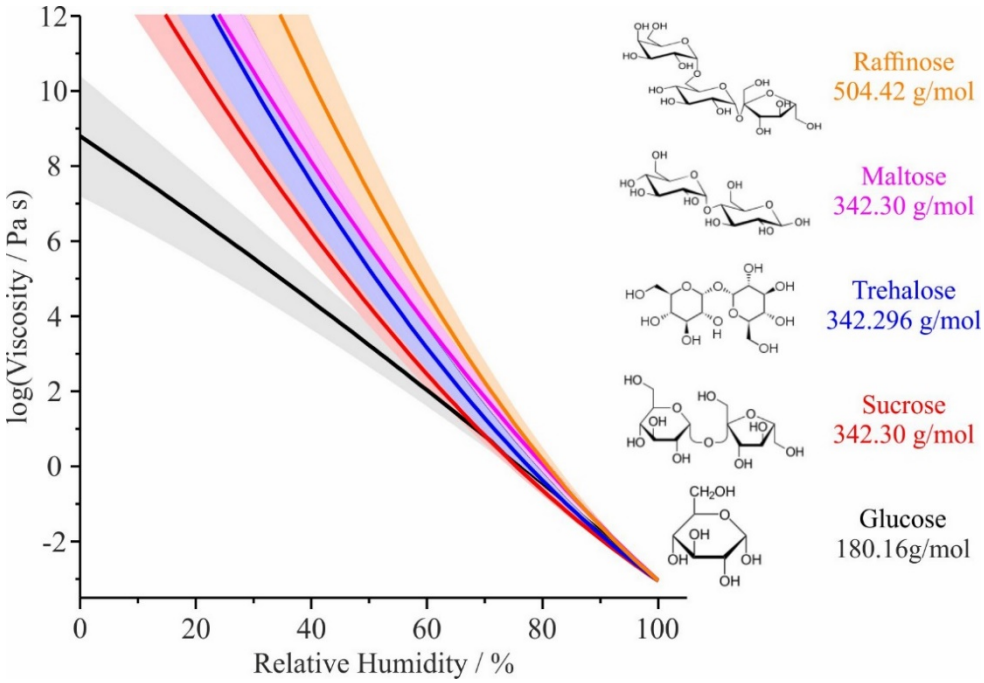


FIGURE 8.6: COMPARISON OF VISCOSITY OF DIFFERENT SYSTEMS. Aqueous glucose (black), aqueous sucrose (red) aqueous trehalose (blue), aqueous maltose (pink), and aqueous raffinose (orange) [285]. Figure is redrawn from Song et. al. [285]

8.7 Molecular dynamics

8.7.1 Force Fields

Water was represented by the TIP4P/2005 [417] potential, due to its accuracy in reproducing the experimental phase and self diffusion characteristics. Sucrose was represented by a modified version of the GROMOS 54a7 force field [418], with an expanded range of atom types. Both the force field and the initial pdb all-atom coordinates were acquired from the automated topology builder (ATB) database [419] (further details on the generation [420] and validation [421] of coordinates, partial charges and force fields by ATB has been detailed extensively in the literature).

Initial coordinates were generated using the Packmol [422] program, which randomly places set numbers of molecules into three dimensional space, allowing tight control over the solute mole fractions in the generated simulation boxes, whilst not biasing the simulations to one area of configuration space. Constraints were inserted such that no two molecules were placed within $3\,\text{Å}$ of each other, and the input random seed was continuously replaced using the bash `$RANDOM` global variable.

The Lincs algorithm [423] was used to constrain all bonds, to an order of four in the constraint coupling matrix, with seven iterations in the final step.Electrostatic forces were calculated using the particle mesh Ewald summation [424], and Van der Waals interactions were provided by the twin range cut-offs method, both of which were truncated at $8\,\text{Å}$. The update frequency was every 5

time-steps. The Verlet scheme [425] was used for neighbour searching across the periodic boundary conditions in three dimensions. Velocities were generated using a Maxwell-Boltzmann distribution at $300\,\mathrm{K}$, with the random seed continuously changed.

8.7.2 Equilibration

The starting coordinates produced by Packmol [422] are not suitable for MD simulations immediately. The configuration must be energy minimised using the steepest descent method to allow bonds and angles to satisfy the constraints of the topology file. The minimisation was conducted with an initial step size of $0.001\,\mathrm{nm}$ until the maximum force was below $50\,\mathrm{kJ\,mol^{-1}\,nm^{-1}}$. After this, approximately $500\,\mathrm{ps}$ of equilibration was conducted with the standard GROMACS MD integrator and thermodynamic ensemble produced as described in the main text.

8.7.3 Mean squared displacement data collection

Initial coordinates were generated from the output frame of the equilibration trajectory within the regime where the total energy was stable. During the MD integration, the system was propagated in the NpT ensemble at $300\,\mathrm{K}$ and $101\,\mathrm{kPa}$, using a velocity rescaling thermostat [233] and the Parrinello-Rahman barostat [426], respectively. Initial velocities were randomly generated to satisfy a Maxwell-Boltzmann distribution at $300\,\mathrm{K}$ in each case. All simulations described were conducted using GPU acceleration, with the MD package GROMACS (version 5.0.6) [427], running on the Blue Crystal 3 high performance computing cluster at the University of Bristol.

Due to the extremely kinetically limited state of the aqueous-sucrose system, it is necessary to simulate dynamics for very long periods of time, relative to, for example, the timescale of molecular vibration or rotation. This maximises the probability that the initial conditions are overcome, and that the constituent molecules decorrelated from their initial conditions as the simulation proceeds. In each of the trajectories of caged water, dynamics were computed for $1\,\mu\mathrm{s}$, and it was found that, when averaged over all nine trajectories, the mean squared displacements of the water molecules converged to a diffusive dependence on simulation time, t (figure 8.7).

$$\left\langle r^2 \right\rangle = 2D_{\mathrm{w,org}}t \tag{8.5}$$

The net displacement achieved by a random walk in three-dimensions will be less than the path length taken between the initial and final positions, a consequence of the fact that the translational motion of water through such a lattice is not Brownian. All nine trajectories contain segments during which the water is travelling perpendicularly, or even backwards, relative to its net displacement. It may be the case that the decorrelation of the water velocity is limited by the rearrangement of the sucrose, as well as the thermodynamic barrier that must be overcome to hop to the nearest available cavity, rather than the dynamics of a more typical solvation shell in an aqueous environment. Therefore, it is desirable to calculate the magnitude of the displacement to the current position, $\left|\left\langle r^2 \right\rangle\right|$, at every time-step, rather than cumulatively sum the path length that takes into account

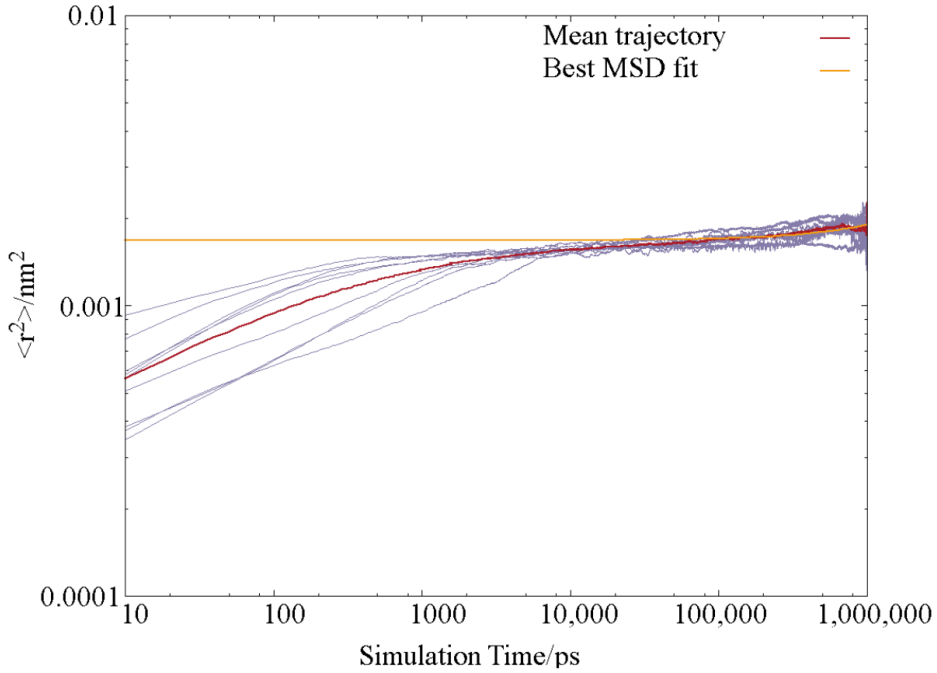


FIGURE 8.7: SINGLE PARTICLE MEAN SQUARED DISPLACEMENT. Shown are the $\langle r^2 \rangle$ of the nine trajectories (purple), along with best fit of the mean path (red) to equation 8.5 (yellow, including constant at $t = 0$)

every intermediate 'jump'. Additionally, removal from the calculation of the centre of mass motion of the water molecule ensures that any atomically resolved vibrational motion does not contribute to the calculation of net displacement. The value stated in the manuscript and illustrated in figure 3.1 panel (c) was calculated under these conditions.

These MD simulations were not designed to capture diffusion against a chemical potential gradient, which is what is induced and probed during the optical tweezers measurements reported above. Instead we wished to treat the water as a tracer particle moving stochastically through the matrix, subject to a small localization uncertainty arising from motion between adjacent frames [428]. This is the origin of the offset at $t = 0$ in the figure. The per simulation D values are presented in table 8.2.

A literature review was conducted to investigate whether any corrections or different functional forms of mean squared displacement needed to be fit. A recent publication [429] by Alcazar-Cano and Delgado-Buscalioni has suggested that in systems where the diffusing 'tracer' particles are trapped in the manner described here, and cannot freely move through channels, it may be more appropriate to fit $\langle r^2 \rangle$ to a subdiffusive dependence, namely,

$$\langle r^2 \rangle \sim D_{\mathrm{w,org}} t^{\alpha} \tag{8.6}$$

where $\alpha \to 0$ as the proportion of particles that are trapped approaches 100 %. Similar physics was described by Zwanzig [1] earlier, in 1988, although it was incorporated into the mathematics by

correcting D, rather than t. Specifically, Zwanzig considered a so-called 'rough potential' where the particle under consideration must traverse a landscape of many nearly degenerate local minima. In that case, the observed D is smaller than the effective D by a factor ϵ that accounts for the 'roughness' of the free energy landscape.

$$D_{\mathrm{w,org}} = D_{\mathrm{effective}} \exp\left(-\frac{\epsilon}{kT}\right)^2 \tag{8.7}$$

These phenomena are often observed in conjunction with stretched exponential relaxation in sugar solutions [282, 430], which the radius curves in figure 3.1 panel (a) also exhibit.

8.7.4 Calculation free volume

Three repeat trajectories of $10\,\mathrm{ns}$ length were conducted for glucose and raffinose, once again generated with randomly placed and oriented molecules via packmol and containing a single water. To determine the sucrose packing efficiency, truncated trajectories $10\,\mathrm{ns}$ long were extracted from the microsecond trajectories and subject to the same analysis, using the GROMACS 'free-volume' program, which attempts to insert 'dummy' probe particles into the box. The free volume calculated is the total volume of the successful insertions.

8.8 Markov state modelling

This section of the supplementary material has been incorporated into chapter 3.

8.9 The comparison of diffusion coefficient of aqueous-sucrose system

For diffusivity research, sucrose is a representative organic system because many researches have used sucrose. The comparison of diffusion curve of sucrose as a function of water activity is shown in figure 8.8. The diffusion constants of water in aqueous-sucrose system in Figure S11 are measured and simulated with inferred by different strategies using AOT with Fi-Pad model, EDB with ETH model, and isotopic exchange. These three lines show good agreement until water activity 0.35. The diffusion constants of water at 0.2 water activity by Price et al. [121] and this research differ by one order of magnitude.

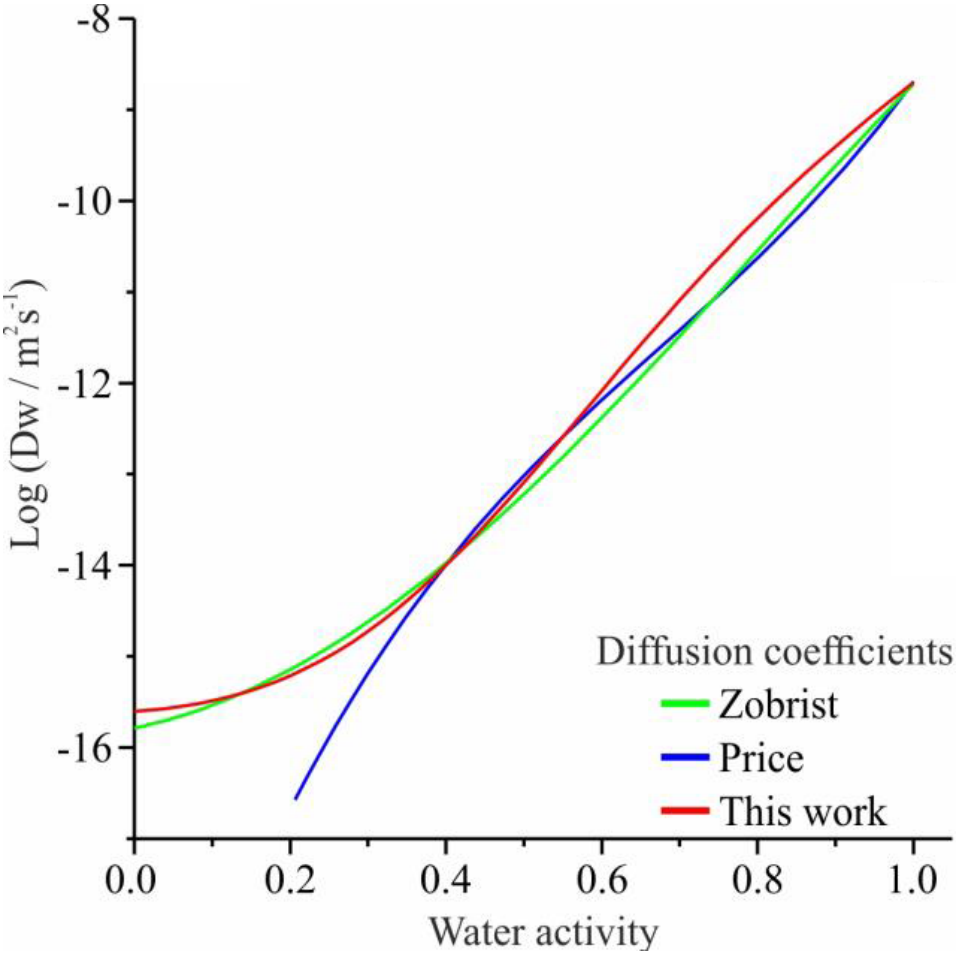


FIGURE 8.8: THE DIFFUSION CURVE OF WATER IN A SUCROSE SYSTEM AS A FUNCTION OF WATER ACTIVITY. The green line is from Zobrist et al. [431], the blue line from Price et al. [121], and the red line is the experimental data which is simulated by Fi-PaD model.

RH transition (%)	Particle size (nm)	Characteristic timescale (s)	RH transition (%)	Particle size (nm)	Characteristic timescale (s)
Aqueous-Glucose system			**Aqueous-Maltose system**		
50-30	3707	27	50-30	368	3721
–	3692	27	–	514	4114
–	3871	29	–	534	4329
–	4861	33	–	549	5131
50-20	3707	49	30-5	3664	3289
–	3692	51	–	2624	3446
–	4861	92	–	4062	3745
30-5	3377	860	–	4845	4197
–	4053	1007	–	3996	4355
–	3871	1009	–	5755	4662
–	4412	1084	–	1641	4130
–	4988	1289	30-50	436	3265
–	5450	1326	–	423	3268
30-50	3707	49	–	441	4101
–	3692	51	–	597	5012
–	4861	63	–	617	5430
20-50	3707	95	5-30	960	3446
5-30	3377	140	–	1490	4197
–	4412	190	–	1250	4662
–	5450	276	–	1230	4145
–	–	–	–	578	4130
Aqueous-Raffinose system			**Aqueous-Sucrose system**		
50-30	3332	188	50-10	3052	397
–	3414	221	–	3307	535
–	3967	250	–	4092	883
–	4110	265	–	4013	1019
–	5005	298	–	4954	1076
–	4924	313	50-20	3796	244
50-20	3260	153	–	4189	363
–	3332	254	–	4783	540

Continued on next page

Table 8.0 – continued from previous page

RH transition (%)	Particle size (nm)	Characteristic timescale (s)	RH transition (%)	Particle size (nm)	Characteristic timescale (s)
–	3410	288	–	4942	559
–	3966	284	50-30	3682	48
–	3606	260	–	4016	59
–	3779	300	–	4894	65
–	4105	320	–	4923	60
–	5008	390	–	5745	85
50-10	3260	178	30-20	3068	668
–	3410	352	–	3123	847
–	3719	700	–	3651	1502
–	5008	1010	–	4382	1748
30-10	3551	1210	–	4758	2274
–	5876	1906	30-10	3297	1243
30-5	3551	895	–	3654	1786
–	3735	995	–	3655	1901
–	3952	1447	–	3652	1924
–	4343	1578	–	4020	2141
–	5895	2101	–	4017	2191
30-50	3332	482	–	4308	2277
–	3414	495	–	3891	2379
–	3967	547	–	5727	3582
–	4110	738	–	4824	4374
–	4924	617	30-5	3246	1321
–	5005	730	–	3657	1804
20-50	3262	432	–	3809	2035
–	3332	351	–	3754	2061
–	3414	440	–	4987	3584
–	3606	476	–	5163	3738
–	3967	399	–	5266	3763
–	5005	516	30-50	3426	57
5-30	3551	610	–	3829	77
–	3735	514	–	4042	79

Continued on next page

Table 8.0 – continued from previous page

RH transition (%)	Particle size (nm)	Characteristic timescale (s)	RH transition (%)	Particle size (nm)	Characteristic timescale (s)
–	3952	852	–	4962	112
–	4343	646	5-30	3252	1776
–	5895	926	–	3645	2059
–	–	–	–	3563	2395
–	–	–	–	3633	2566
–	–	–	–	3983	2763
–	–	–	–	4716	3056
–	–	–	–	4912	3294
–	–	–	–	5912	4112
Aqueous-Raffinose system			**Aqueous-Sucrose system**		
50-30	3265	98	50-30	3590	35
–	3268	101	–	3588	40
–	4101	157	–	3746	32
–	5430	231	–	4432	43
–	5012	204	–	4463	47
50-20	3265	221	–	5448	53
–	3268	194	30-50	3588	81
–	4101	184	–	3746	52
–	5012	244	–	4432	117
–	5430	233	–	4463	86
30-5	3090	1019	–	5448	76
–	3190	1172	–	–	–
–	3285	1155	–	–	–
–	3310	1184	–	–	–
–	4040	1342	–	–	–
–	4825	1479	–	–	–
–	5168	1559	–	–	–
–	5118	1444	–	–	–
30-50	3265	436	–	–	–
–	3268	423	–	–	–
–	4101	441	–	–	–

Continued on next page

Table 8.0 – continued from previous page

RH transition (%)	Particle size (nm)	Characteristic timescale (s)	RH transition (%)	Particle size (nm)	Characteristic timescale (s)
–	5430	617	–	–	–
–	5012	597	–	–	–
20-50	3265	181	–	–	–
5-30	4040	726	–	–	–
–	3190	515	–	–	–
–	3090	592	–	–	–
–	5168	846	–	–	–
–	3853	589	–	–	–
–	4897	719	–	–	–
–	5118	981	–	–	–
–	3310	596	–	–	–

TABLE 8.1: THE CHARACTERISTIC TIMESCALE OF WATER TRANSPORT DETERMINED EXPERIMENTALLY FOR SIX BINARY MIXTURES. This table provides all data points for water transport kinetics in figure 8.2. Particle size is direct measurement data in AOT and fit by LARA. The characteristic timescale is fit by KWW function. Error representing a variation of $\beta \pm 0.1$.

Trajectory no.	D ($m^2 s^{-1}$)	Uncertainty ($m^2 s^{-1}$)
1	6.454E-17	9.34E-17
2	4.848E-17	1.755E-16
3	6.075E-17	1.444E-16
4	8.922E-18	3.448E-17
5	1.004E-18	1.88E-16
6	8.775E-17	2.218E-17
7	3.541E-17	8.571E-17
8	3.829E-17	1.61E-17
9	7.269E-17	5.83E-17

TABLE 8.2: DIFFUSION CONSTANTS FROM MD TRAJECTORIES. The best fit diffusion coefficients to the nine 1 µs trajectories (purple lines in figure 8.7) and associated uncertainties.

System	Paramaterization ($D_w(a_w)$ ($m^2\,s^{-1}$))
Aqueous-Levoglucosan	$\log(D_w(a_w)) = -14.091 + (2.192 \times a_w) + \left(-21.269 \times (a_w)^2\right) + \left(81.025 \times (a_w)^3\right) + \left(-84.660 \times (a_w)^4\right) + \left(28.081 \times (a_w)^5\right)$
Aqueous-glucose	$\log(D_w(a_w)) = -15.047 + (0.963 \times a_w) + \left(-0.186 \times (a_w)^2\right) + \left(34.825 \times (a_w)^3\right) + \left(-47.724 \times (a_w)^4\right) + \left(18.472 \times (a_w)^5\right)$
Aqueous-sucrose	$\log(D_w(a_w)) = -15.613 + (1.262 \times a_w) + \left(-3.476 \times (a_w)^2\right) + \left(46.468 \times (a_w)^3\right) + \left(-60.030 \times (a_w)^4\right) + \left(22.691 \times (a_w)^5\right)$
Aqueous-Trehalose	$\log(D_w(a_w)) = -15.503 + (1.061 \times a_w) + \left(0.642 \times (a_w)^2\right) + \left(34.712 \times (a_w)^3\right) + \left(-48.590 \times (a_w)^4\right) + \left(18.981 \times (a_w)^5\right)$
Aqueous-Maltose	$\log(D_w(a_w)) = -15.857 + (1.269 \times a_w) + \left(-6.922 \times (a_w)^2\right) + \left(58.525 \times (a_w)^3\right) + \left(-72.854 \times (a_w)^4\right) + \left(27.141 \times (a_w)^5\right)$
Aqueous-Raffinose	$\log(D_w(a_w)) = -15.393 + (1.113 \times a_w) + \left(2.180 \times (a_w)^2\right) + \left(29.194 \times (a_w)^3\right) + \left(-42.838 \times (a_w)^4\right) + \left(17.051 \times (a_w)^5\right)$

TABLE 8.3: SIX BINARY SYSTEM DIFFUSION COEFFICIENTS. The best fit diffusion coefficients of six binary systems in figure 3.1 panel (c).

CHAPTER 9

MARKOV STATE MODEL OPTIMISATION

T(n)	Name	SMSE	MSLL
$\log(n)$	Exponential	0.0012	-3.9963
	Matérn 3-2	0.0010	-4.1712
	Matérn 5-2	0.0007	-4.2369
	Gaussian	0.0011	-4.0892
$I(n)$	Exponential	0.0027	-2.9733
	Matérn 3-2	0.0025	-3.4218
	Matérn 5-2	0.0023	-3.8172
	Gaussian	0.0032	-4.1239

TABLE 9.1: GAUSSIAN PROCESS MODEL SELECTION METRICS OF THE RESPONSE SURFACE OF ALANINE DIPEPTIDE. Standardised mean square error (SMSE) and mean standardised log loss (MSLL) for GP models of the response surface of MSMs for alanine dipeptide, using different transformations of n, $T(n)$, and different kernels. Each GP model used a mean prior of zero, and all other parameters were estimated by maximizing the marginal likelihood. All values were calculated using 10-fold cross-validation.

CHAPTER 10

METASTABLE STATE SELECTION FOR HIDDEN MARKOV MODELS

10.1 Prinz potential

The Prinz potential is given by:

$$V(x) = 4\left(x^8 + 0.8\exp\left(-80x^2\right) + 0.2\exp\left(-80(x-0.5)^2\right) + 0.5\exp\left(-40(x+0.5)^2\right)\right). \tag{10.1}$$

Exact eigenvalues and trajectories of simulated Brownian motion were calculated using code from MSMBuilder (version 3.9.0) [406] . The first 7 relaxation processes are given in table 10.1.

100 trajectories of Brownian dynamics were simulated by the following stochastic differential equation:

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = -\frac{\mathrm{d}V(x)}{\mathrm{d}x} + \sqrt{2D}\cdot R(t) \tag{10.2}$$

with $D = 1000$, and $R \sim \mathcal{N}(0,1)$, $\mathrm{Cov}\left[R(t), R(t')\right] = \delta_{t,t'}$. The time-step used was $\Delta t = 0.001$. Each trajectory was initiated from a random draw of the stationary distribution and was twice the longest

Process, i	t_i
2	844.4
3	125.5
4	64.3
5	11.9
6	10.3
7	7.3
8	6.7

TABLE 10.1: RELAXATION TIMESCALES OF THE PRINZ POTENTIAL. All values are given in units of the time step $\Delta t = 0.001$

relaxation process timescale, i.e. $2 \times 844 = 1688$ time-steps long. The trajectories were clustered into $n = \left\lfloor \sqrt{100 \times 1688} \right\rfloor = 410$ discrete states using k-means clustering [150]. This number of states was based on the heuristic described in reference [246].

10.2 Membership matrix errors

The conditional probabilities $\mathbb{P}(h = j | s = i)$ can be calculated via the γ auxiliary variable from the Baum-Welch algorithm [170, 258], or from the membership matrix after taking into account the largest strongly connected subset of hidden states. Figure 10.1 shows the difference between these two methods for a HMM with $g = 10$ hidden states. The error plotted in panel (a) (blue discs) is given by:

$$\mathrm{err}(s_i) = \sum_j \bar{\gamma}_j(s_i) - M_{ij}, \tag{10.3}$$

and is plotted as a function of the stationary distribution of the observed states π_{s_i}. The green discs are the stationary distribution of states with zero error $(\mathrm{err} < 10^{-10})$. $\bar{\gamma}$ is the value of $\gamma_j(s_i)$ averaged over all the instances of a s_i in a trajectory. As the size of the error decreases with frequency of the observed state, it can be concluded that the difference between the two methods arises from sampling error. Panel (b) shows the value of the error for just one hidden state (state $j = 3$) with the emission distribution for comparison.

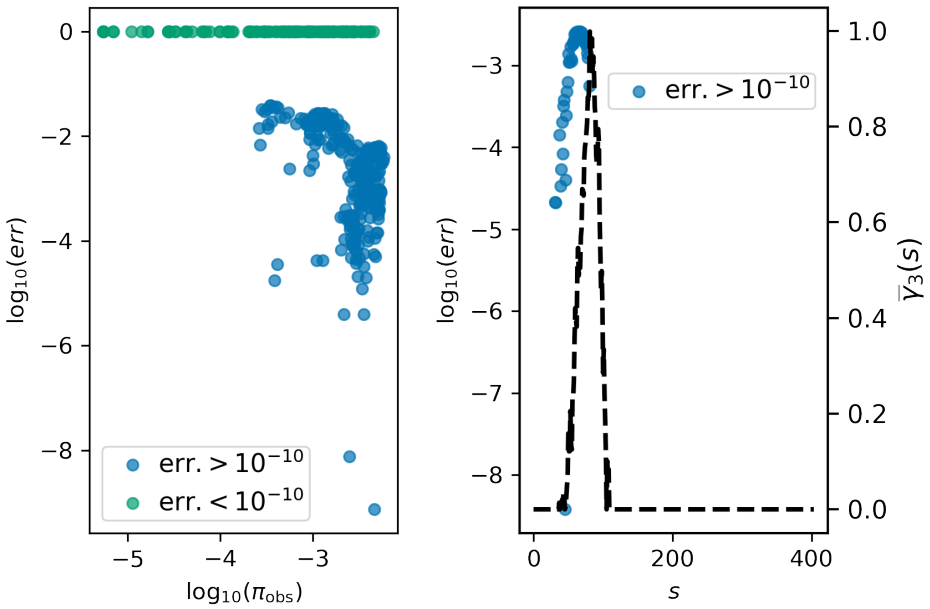


FIGURE 10.1: ERRORS IN THE MEMBERSHIP MATRIX. Panel (a) shows the error from equation 10.3 (blue discs) as a function of the observed stationary distribution, $\boldsymbol{\pi}_{\text{obs}}$. The green discs show the observed states with no difference in value of γ or $\mathbf{M}$. Panel (b) shows the error for hidden state 3 (blue discs) overlaid on the value of $\gamma_3(s)$ as a function of s, the observed states.

CHAPTER 11

AROMATIC AMINE DEHYROGENASE

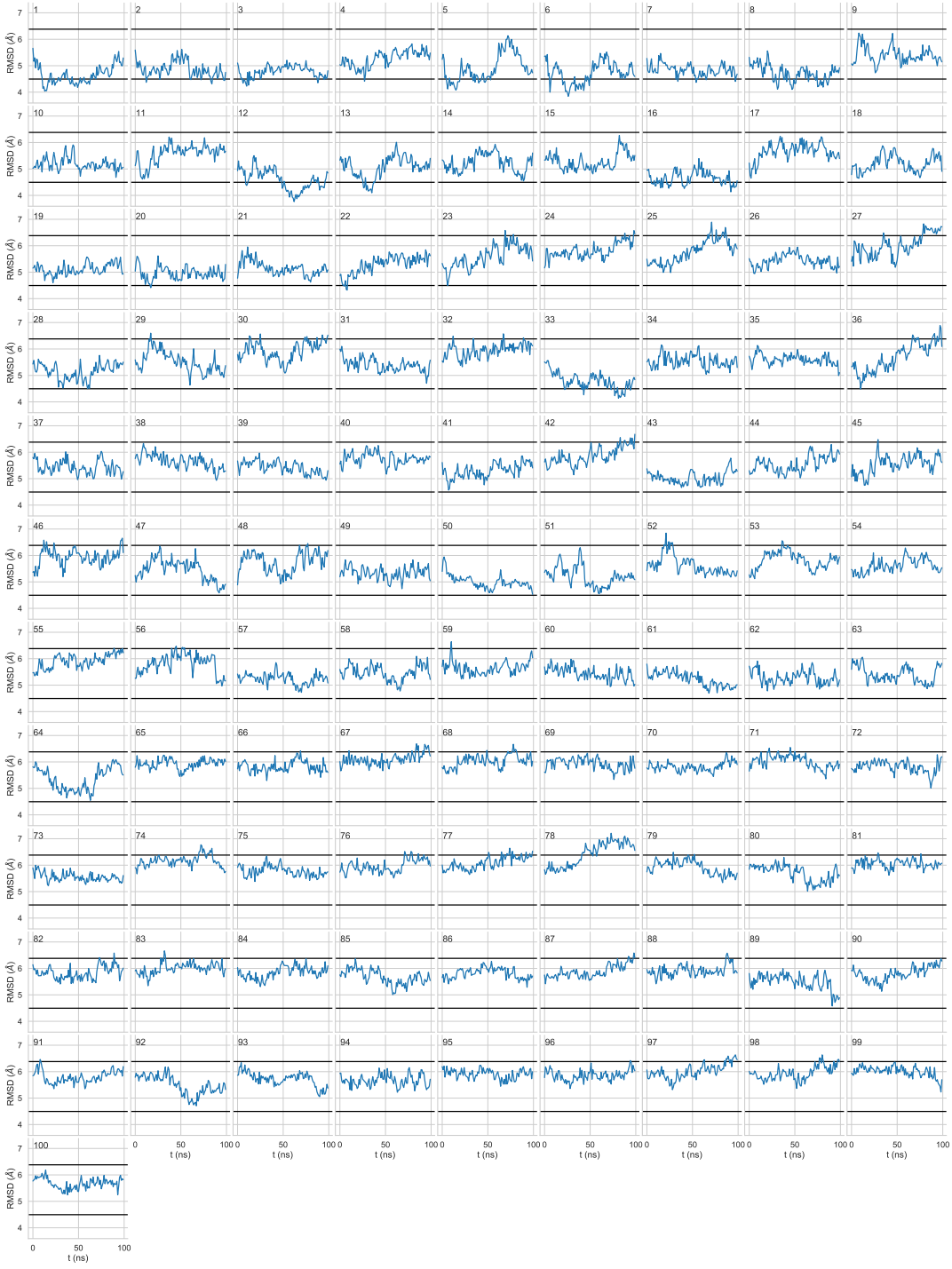


FIGURE 11.1: RMSD OF THE α-CARBON ATOMS OF AADH RELATIVE TO THE CRYSTAL STRUCTURE. Each panel is a single trajectory, blue lines are the RMSD, horizontal lines are the 2.5 % and 97.5 % quantiles (4.5 Å and 6.4 Å, respectively) taken across all trajectories.

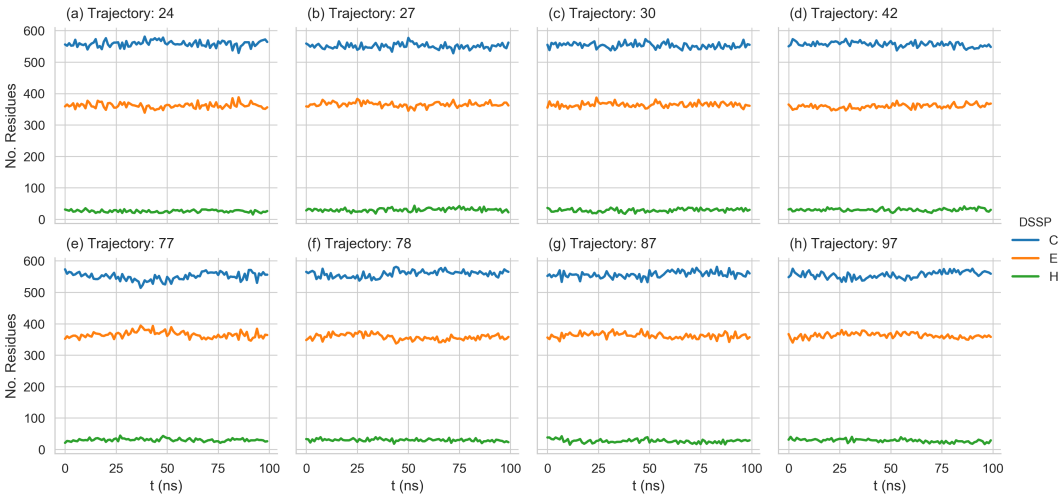


FIGURE 11.2: SECONDARY STRUCTURE COMPOSITION OF TRAJECTORIES 24, 27, 30, 42, 78, 87 AND 97 AS A FUNCTION OF TIME. The number of residues in each simplified secondary structure class [432] are shown: 'H' (green) refers to alpha helix, 3- and 5-helices; 'E' (orange) refers to residues in beta-bridges or beta ladder; 'C' (blue) refers to turns, bends and all irregular elements.

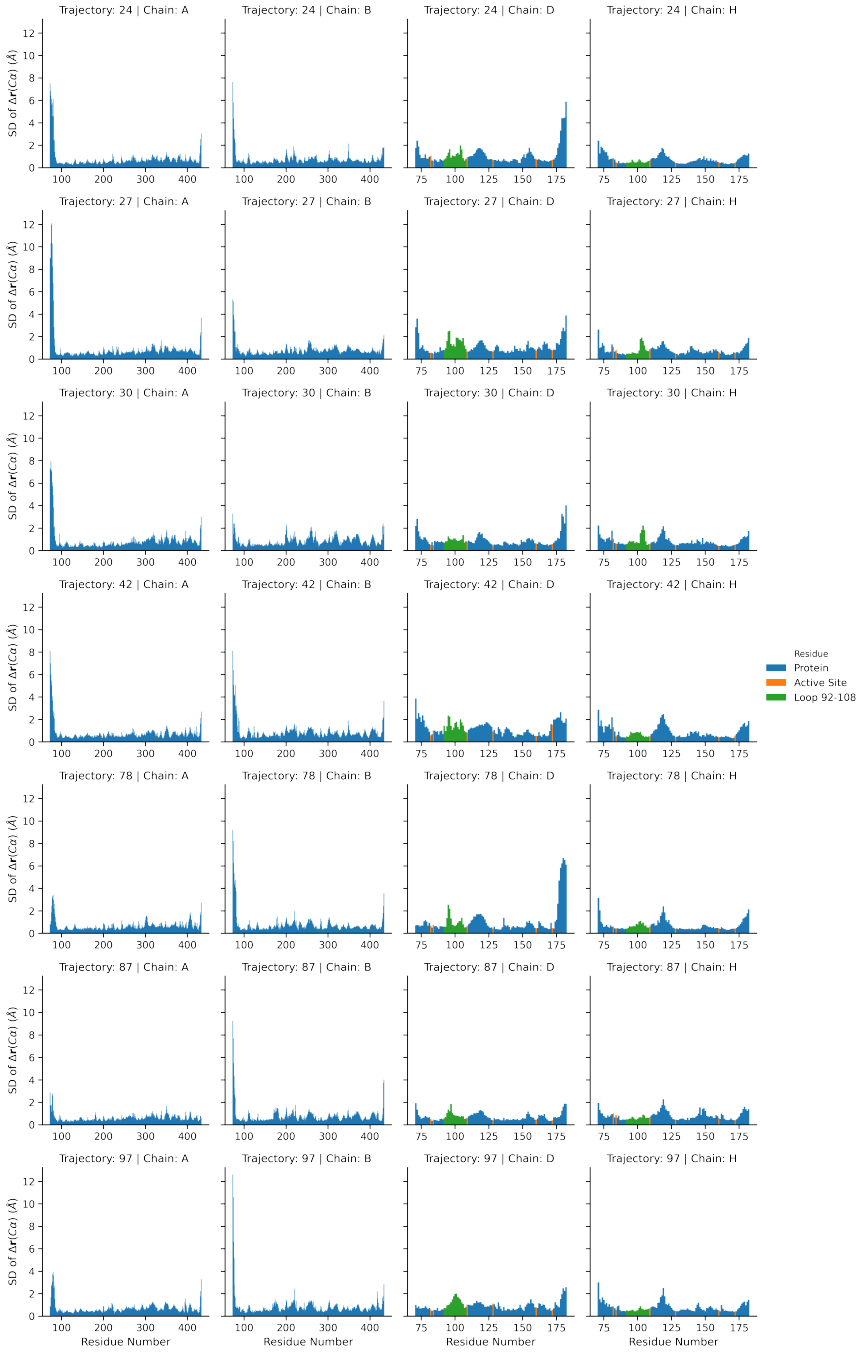


FIGURE 11.3: FLUCTUATIONS IN THE DEVIATION OF RESIDUES OF TRAJECTORIES 24, 27, 30, 42, 78, 87 AND 97. Each row corresponds to a different trajectory, each column to a different chain. 100 regularly spaced snapshots were taken from each trajectory and aligned to the crystal structure along with α-carbon atoms. The standard deviation of the deviation of the α-carbon atoms from the crystal structure, $\Delta\mathbf{r} = \mathbf{r} - \mathbf{r}_{\text{crystal}}$ is plotted for each residue.

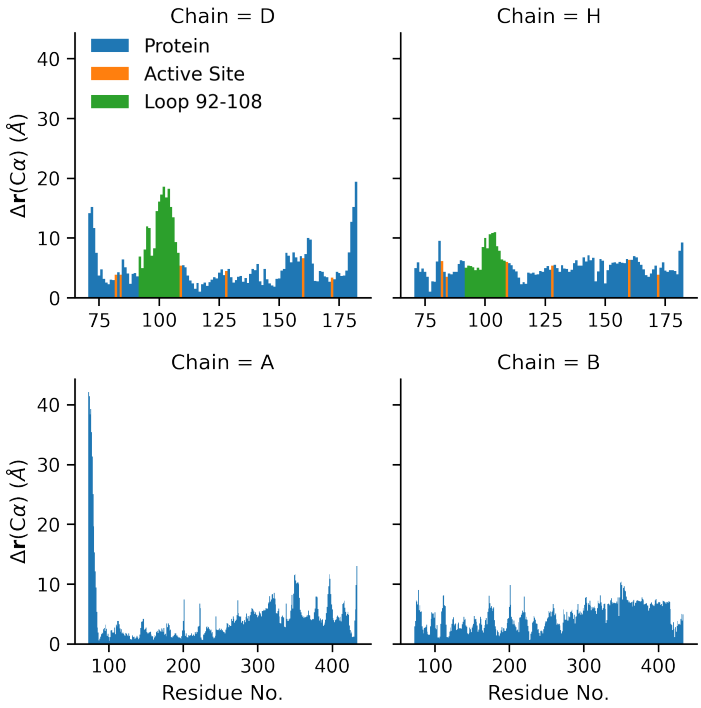


FIGURE 11.4: α-CARBON DEVIATION OF RESIDUES FOR SEED TRAJECTORY AT 95 ns. Each panel depicts a separate chain of the structure from a snapshot of the seed trajectory at 95 ns. The RMSD is 6.2 Å, the highest value reached, see figure 6.4 panel (a). In chains D and H the six residues of the active site are shown in orange and residues 92–108 are shown in green. The conformational change of loop 92–108 is shown in figure 6.5

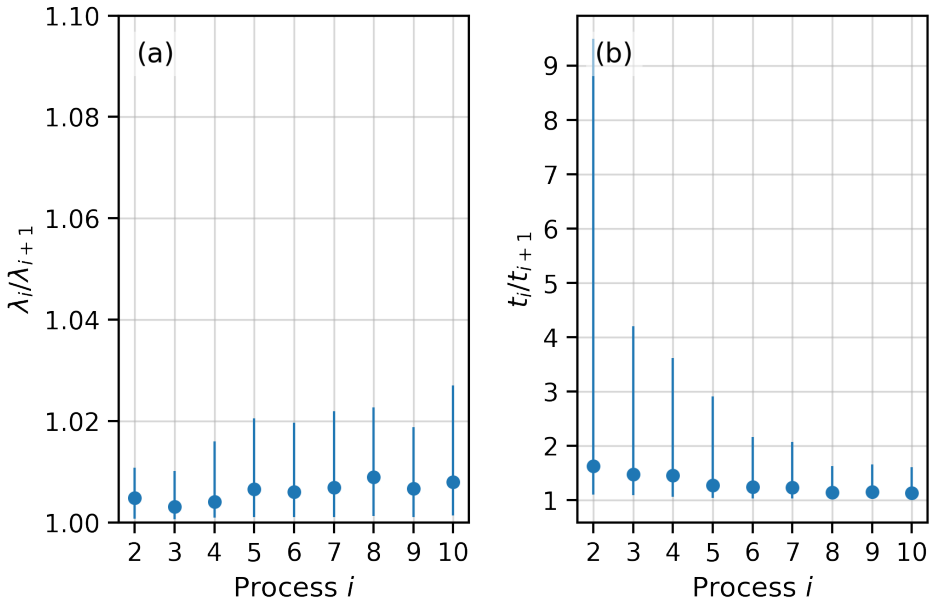


FIGURE 11.5: THE RATIO OF SUCCESSIVE EIGENVALUES AND IMPLIED TIMESCALES OF THE SENSITIVITY REFERENCE MSM. Panel (a) shows the ratio of successive eigenvalues and panel (b) the implied timescales the sensitivity case of the reference MSM with: $\tau_{\mathrm{M}} = 2\,\mathrm{ns}$, TICA lag time of $\tau = 10\,\mathrm{ns}$, 95 % of the kinetic variance/$m = 8$ TICA components retained, and $n = 316$ microstates. Parameters were estimated using MCMC with 1000 posterior samples, the blue dots and error bars are the mean and 95 % credible intervals respectively.

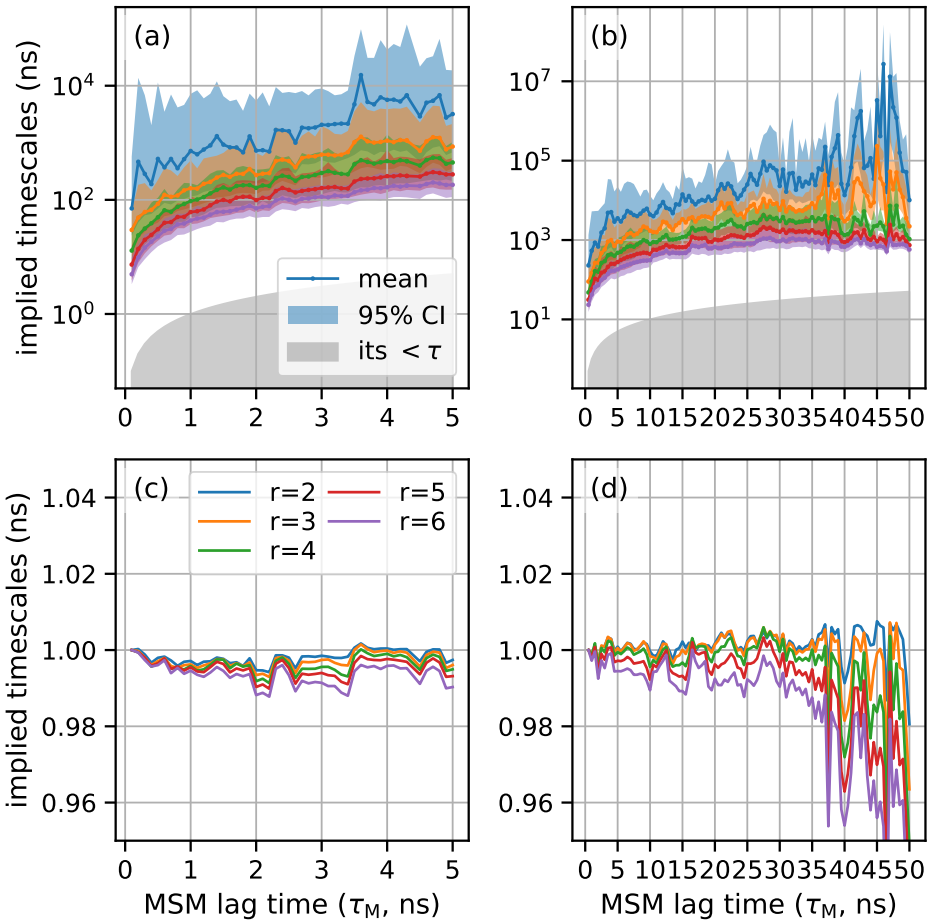


FIGURE 11.6: THE IMPLIED TIMESCALES AND VAMP-2 SCORES OF THE SENSITIVITY REFERENCE MSM. Panels (a) and (b) show the implied timescales, and panels (c) and (d) shows the relative VAMP-2 scores for the reference MSM with: $\tau_M = 2\,\text{ns}$, TICA lag time of $\tau = 10\,\text{ns}$, 95 % of the kinetic variance/$m = 10$ TICA components retained, and $n = 316$ microstates. Panel (a) shows the first five implied timescales for $\tau_M = 0.1\,\text{ns}$–$5\,\text{ns}$, panel (b) shows the first five implied timescales for $\tau_M = 0.1\,\text{ns}50\,\text{ns}$. The solid lines and coloured shaded areas are the mean and 95 % credible intervals respectively, estimated using MCMC with 500 posterior samples. The grey shaded area is the region for which the implied timescales are smaller than the lag time. Panel (c) and (d) show the VAMP-2 scores, scored on the first 2 to 5 eigenvalues for the same ranges. The VAMP-2 scores are indexed to their value at $\tau_M = 0.1\,\text{ns}$. The colour coding is consistent between the implied timescale plots ((a) and (b)) and VAMP-2 plots ((c) and (d)). e.g. the blue line in (c) and (d) is the VAMP-2 score with two eigenvalues ($r = 2$) while in (a) and (b) blue is the second implied timescale, t_2.

$T(\tau)$	$T(m)$	$T(n)$	Kernel	MSLL	SMSE	Rank (MSLL)	Rank (SMSE)	Rank (Total)
$I(\tau)$	$I(m)$	$I(n)$	Exponential	-0.1298	0.3087	1.0	1.0	1.0
		$\log(n)$	Exponential	0.0050	0.2964	-	-	-
	$\log(m)$	$I(n)$	Exponential	0.0521	0.3118	-	-	-
		$\log(n)$	Exponential	0.5633	0.3815	-	-	-
$\log(\tau)$	$I(m)$	$I(n)$	Exponential	0.1967	0.3436	-	-	-
		$\log(n)$	Exponential	0.4959	0.3231	-	-	-
	$\log(m)$	$I(n)$	Exponential	0.5128	0.4365	-	-	-
		$\log(n)$	Exponential	1.0267	0.4201	-	-	-
$I(\tau)$	$I(m)$	$I(n)$	Matérn 3-2	1.5680	0.2893	-	-	-
		$\log(n)$	Matérn 3-2	1.9193	0.2960	-	-	-
	$\log(m)$	$I(n)$	Matérn 3-2	3.1385	0.2775	-	-	-
		$\log(n)$	Matérn 3-2	2.0358	0.2818	-	-	-
$\log(\tau)$	$I(m)$	$I(n)$	Matérn 3-2	6.9015	0.3203	-	-	-
		$\log(n)$	Matérn 3-2	7.7182	0.3406	-	-	-
	$\log(m)$	$I(n)$	Matérn 3-2	7.9209	0.3257	-	-	-
		$\log(n)$	Matérn 3-2	3.0002	0.3472	-	-	-
$I(\tau)$	$I(m)$	$I(n)$	Matérn 5-2	3.6517	0.3029	-	-	-
		$\log(n)$	Matérn 5-2	3.8316	0.3090	-	-	-
	$\log(m)$	$I(n)$	Matérn 5-2	8.6574	0.2991	-	-	-
		$\log(n)$	Matérn 5-2	3.7354	0.3238	-	-	-
$\log(\tau)$	$I(m)$	$I(n)$	Matérn 5-2	8.9207	0.3679	-	-	-
		$\log(n)$	Matérn 5-2	11.8753	0.4064	-	-	-
	$\log(m)$	$I(n)$	Matérn 5-2	12.7637	0.3722	-	-	-
		$\log(n)$	Matérn 5-2	13.6735	0.3475	-	-	-
$I(\tau)$	$I(m)$	$I(n)$	Gaussian	inf	inf	-	-	-
		$\log(n)$	Gaussian	9.3618	0.3256	-	-	-
	$\log(m)$	$I(n)$	Gaussian	5.9123	0.3022	-	-	-
		$\log(n)$	Gaussian	inf	inf	-	-	-
$\log(\tau)$	$I(m)$	$I(n)$	Gaussian	17.4786	0.4551	-	-	-
		$\log(n)$	Gaussian	16.7568	0.3556	-	-	-
	$\log(m)$	$I(n)$	Gaussian	16.7412	0.5026	-	-	-
		$\log(n)$	Gaussian	24.3199	0.4986	-	-	-

TABLE 11.1: GAUSSIAN PROCESS MODEL SELECTION METRICS FOR THE RESPONSE SURFACE OF AADH USING ALL RANDOMLY SAMPLED HYPERPARAMETER TRIALS, $\mathscr{D}_{361}$. The mean standardised log loss (MSLL) and standardised mean square error (SMSE) where calculated using 10 fold cross validation. Only those models which had both $\text{MSLL} < 0$ and $\text{SMSE} < 1$ were ranked. The total rank is calculated as rank of $\left(R^2_{\text{MSLL}} + R^2_{\text{SMSE}}\right)^{1/2}$. Where the overall rank was tied, the first model appearing in the table was ranked higher.

$T(\tau)$	$T(m)$	$T(n)$	Kernel	MSLL	SMSE	Rank (MSLL)	Rank (SMSE)	Rank (Total)
$I(\tau)$	$I(m)$	$I(n)$	Exponential	-0.3928	0.3412	10.0	14.0	13.0
		$\log(n)$	Exponential	-0.2456	0.3443	15.0	15.0	16.0
	$\log(m)$	$I(n)$	Exponential	-0.6484	0.3169	6.0	12.0	7.0
		$\log(n)$	Exponential	-0.5585	0.3487	7.0	17.0	14.0
$\log(\tau)$	$I(m)$	$I(n)$	Exponential	0.0598	0.3483	-	-	-
		$\log(n)$	Exponential	-0.2195	0.3450	16.0	16.0	17.0
	$\log(m)$	$I(n)$	Exponential	-0.3524	0.3084	13.0	9.0	10.0
		$\log(n)$	Exponential	-0.3944	0.3379	9.0	13.0	11.0
$I(\tau)$	$I(m)$	$I(n)$	Matérn 3-2	-0.3807	0.3167	12.0	11.0	12.0
		$\log(n)$	Matérn 3-2	-0.2744	0.3053	14.0	7.0	9.0
	$\log(m)$	$I(n)$	Matérn 3-2	-0.8769	0.2779	1.0	4.0	3.0
		$\log(n)$	Matérn 3-2	-0.7438	0.2785	5.0	5.0	5.0
$\log(\tau)$	$I(m)$	$I(n)$	Matérn 3-2	0.3415	0.3721	-	-	-
		$\log(n)$	Matérn 3-2	-0.2023	0.3892	17.0	18.0	18.0
	$\log(m)$	$I(n)$	Matérn 3-2	-0.4758	0.3033	8.0	6.0	6.0
		$\log(n)$	Matérn 3-2	-0.3892	0.3086	11.0	10.0	8.0
$I(\tau)$	$I(m)$	$I(n)$	Matérn 5-2	0.3362	0.3149	-	-	-
		$\log(n)$	Matérn 5-2	0.9964	0.2712	-	-	-
	$\log(m)$	$I(n)$	Matérn 5-2	-0.8713	0.2685	2.0	2.0	1.0
		$\log(n)$	Matérn 5-2	-0.7508	0.2700	4.0	3.0	4.0
$\log(\tau)$	$I(m)$	$I(n)$	Matérn 5-2	6.4201	0.3503	-	-	-
		$\log(n)$	Matérn 5-2	5.7695	0.3250	-	-	-
	$\log(m)$	$I(n)$	Matérn 5-2	3.9718	0.3153	-	-	-
		$\log(n)$	Matérn 5-2	inf	inf	-	-	-
$I(\tau)$	$I(m)$	$I(n)$	Gaussian	-0.1677	0.3074	18.0	8.0	15.0
		$\log(n)$	Gaussian	1.3068	0.2747	-	-	-
	$\log(m)$	$I(n)$	Gaussian	-0.7884	0.2675	3.0	1.0	2.0
		$\log(n)$	Gaussian	inf	inf	-	-	-
$\log(\tau)$	$I(m)$	$I(n)$	Gaussian	6.8541	0.3472	-	-	-
		$\log(n)$	Gaussian	6.2984	0.3074	-	-	-
	$\log(m)$	$I(n)$	Gaussian	4.8742	0.4157	-	-	-
		$\log(n)$	Gaussian	7.6739	0.5531	-	-	-

TABLE 11.2: GAUSSIAN PROCESS MODEL SELECTION METRICS FOR THE RESPONSE SURFACE OF AADH USING HYPERPARAMETER TRIAL DATA SUBSET 1, $\mathscr{D}^1_{100}$. The mean standardised log loss (MSLL) and standardised mean square error (SMSE) where calculated using 10 fold cross validation. Only those models which had both MSLL < 0 and SMSE < 1 were ranked. The total rank is calculated as rank of $\left(R^2_{\text{MSLL}} + R^2_{\text{SMSE}}\right)^{1/2}$. Where the overall rank was tied, the first model appearing in the table was ranked higher.

$T(\tau)$	$T(m)$	$T(n)$	Kernel	MSLL	SMSE	Rank (MSLL)	Rank (SMSE)	Rank (Total)
$I(\tau)$	$I(m)$	$I(n)$	Exponential	-0.3293	0.4262	13.0	9.0	11.0
		$\log(n)$	Exponential	-0.5222	0.4330	6.0	13.0	8.0
	$\log(m)$	$I(n)$	Exponential	-0.6612	0.3890	1.0	2.0	1.0
		$\log(n)$	Exponential	-0.5843	0.4170	3.0	5.0	3.0
$\log(\tau)$	$I(m)$	$I(n)$	Exponential	-0.3737	0.4590	11.0	15.0	15.0
		$\log(n)$	Exponential	-0.4162	0.4445	8.0	14.0	12.0
	$\log(m)$	$I(n)$	Exponential	-0.3702	0.4281	12.0	10.0	10.0
		$\log(n)$	Exponential	-0.6242	0.4169	2.0	4.0	2.0
$I(\tau)$	$I(m)$	$I(n)$	Matérn 3-2	-0.5737	0.4218	4.0	7.0	5.0
		$\log(n)$	Matérn 3-2	0.1639	0.4432	-	-	-
	$\log(m)$	$I(n)$	Matérn 3-2	-0.5479	0.4282	5.0	11.0	7.0
		$\log(n)$	Matérn 3-2	-0.4595	0.3844	7.0	1.0	4.0
$\log(\tau)$	$I(m)$	$I(n)$	Matérn 3-2	-0.4077	0.4301	9.0	12.0	9.0
		$\log(n)$	Matérn 3-2	inf	inf	-	-	-
	$\log(m)$	$I(n)$	Matérn 3-2	1.0248	0.4609	-	-	-
		$\log(n)$	Matérn 3-2	-0.3902	0.3942	10.0	3.0	6.0
$I(\tau)$	$I(m)$	$I(n)$	Matérn 5-2	1.3964	0.4033	-	-	-
		$\log(n)$	Matérn 5-2	0.3681	0.4475	-	-	-
	$\log(m)$	$I(n)$	Matérn 5-2	-0.1968	0.4237	14.0	8.0	13.0
		$\log(n)$	Matérn 5-2	2.3201	0.4400	-	-	-
$\log(\tau)$	$I(m)$	$I(n)$	Matérn 5-2	3.2132	0.4125	-	-	-
		$\log(n)$	Matérn 5-2	0.5430	0.4473	-	-	-
	$\log(m)$	$I(n)$	Matérn 5-2	1.6455	0.4679	-	-	-
		$\log(n)$	Matérn 5-2	0.7421	0.4378	-	-	-
$I(\tau)$	$I(m)$	$I(n)$	Gaussian	2.3960	0.4042	-	-	-
		$\log(n)$	Gaussian	1.3825	0.4372	-	-	-
	$\log(m)$	$I(n)$	Gaussian	-0.1688	0.4197	15.0	6.0	14.0
		$\log(n)$	Gaussian	3.8725	0.4652	-	-	-
$\log(\tau)$	$I(m)$	$I(n)$	Gaussian	4.1994	0.4244	-	-	-
		$\log(n)$	Gaussian	2.3169	0.4305	-	-	-
	$\log(m)$	$I(n)$	Gaussian	1.7600	0.4764	-	-	-
		$\log(n)$	Gaussian	1.6457	0.4517	-	-	-

TABLE 11.3: GAUSSIAN PROCESS MODEL SELECTION METRICS FOR THE RESPONSE SURFACE OF AADH USING HYPERPARAMETER TRIAL DATA SUBSET 2, $\mathscr{D}^2_{100}$. The mean standardised log loss (MSLL) and standardised mean square error (SMSE) where calculated using 10 fold cross validation. Only those models which had both MSLL < 0 and SMSE < 1 were ranked. The total rank is calculated as rank of $\left(R^2_{\text{MSLL}} + R^2_{\text{SMSE}}\right)^{1/2}$. Where the overall rank was tied, the first model appearing in the table was ranked higher.

$T(\tau)$	$T(m)$	$T(n)$	Kernel	MSLL	SMSE	Rank (MSLL)	Rank (SMSE)	Rank (Total)
$I(\tau)$	$I(m)$	$I(n)$	Exponential	-0.4461	0.5415	14.0	15.0	13.0
		$\log(n)$	Exponential	-0.4350	0.5234	15.0	9.0	10.0
	$\log(m)$	$I(n)$	Exponential	-0.6123	0.5074	6.0	5.0	3.0
		$\log(n)$	Exponential	-0.5378	0.5145	9.0	7.0	5.0
$\log(\tau)$	$I(m)$	$I(n)$	Exponential	-0.3138	0.6006	21.0	25.0	24.0
		$\log(n)$	Exponential	-0.3559	0.5626	20.0	21.0	22.0
	$\log(m)$	$I(n)$	Exponential	-0.4587	0.5449	13.0	17.0	14.0
		$\log(n)$	Exponential	-0.4276	0.5472	18.0	18.0	20.0
$I(\tau)$	$I(m)$	$I(n)$	Matérn 3-2	-0.6003	0.5234	7.0	8.0	4.0
		$\log(n)$	Matérn 3-2	-0.6400	0.5376	5.0	12.0	8.0
	$\log(m)$	$I(n)$	Matérn 3-2	-0.8017	0.4885	2.0	1.0	1.0
		$\log(n)$	Matérn 3-2	-0.8921	0.4892	1.0	2.0	2.0
$\log(\tau)$	$I(m)$	$I(n)$	Matérn 3-2	-0.1904	0.5933	24.0	24.0	25.0
		$\log(n)$	Matérn 3-2	-0.4898	0.5711	11.0	22.0	18.0
	$\log(m)$	$I(n)$	Matérn 3-2	-0.4295	0.5379	17.0	13.0	15.0
		$\log(n)$	Matérn 3-2	-0.6808	0.5350	4.0	11.0	6.0
$I(\tau)$	$I(m)$	$I(n)$	Matérn 5-2	-0.4328	0.5482	16.0	19.0	19.0
		$\log(n)$	Matérn 5-2	-0.5392	0.5262	8.0	10.0	7.0
	$\log(m)$	$I(n)$	Matérn 5-2	-0.7498	0.5506	3.0	20.0	12.0
		$\log(n)$	Matérn 5-2	-0.2438	0.5141	22.0	6.0	16.0
$\log(\tau)$	$I(m)$	$I(n)$	Matérn 5-2	0.4797	0.6354	-	-	-
		$\log(n)$	Matérn 5-2	-0.3604	0.6100	19.0	26.0	23.0
	$\log(m)$	$I(n)$	Matérn 5-2	-0.1492	0.5818	25.0	23.0	26.0
		$\log(n)$	Matérn 5-2	0.1426	0.5263	-	-	-
$I(\tau)$	$I(m)$	$I(n)$	Gaussian	-0.5234	0.5412	10.0	14.0	9.0
		$\log(n)$	Gaussian	-0.4854	0.5436	12.0	16.0	11.0
	$\log(m)$	$I(n)$	Gaussian	inf	inf	-	-	-
		$\log(n)$	Gaussian	-0.1291	0.5002	26.0	3.0	21.0
$\log(\tau)$	$I(m)$	$I(n)$	Gaussian	1.5339	0.5471	-	-	-
		$\log(n)$	Gaussian	-0.0794	0.6378	27.0	27.0	27.0
	$\log(m)$	$I(n)$	Gaussian	-0.2000	0.5068	23.0	4.0	17.0
		$\log(n)$	Gaussian	0.1399	0.4935	-	-	-

TABLE 11.4: GAUSSIAN PROCESS MODEL SELECTION METRICS FOR THE RESPONSE SURFACE OF AADH USING HYPERPARAMETER TRIAL DATA SUBSET 3, $\mathscr{D}^3_{100}$. The mean standardised log loss (MSLL) and standardised mean square error (SMSE) where calculated using 10 fold cross validation. Only those models which had both MSLL < 0 and SMSE < 1 were ranked. The total rank is calculated as rank of $\left(R^2_{\text{MSLL}} + R^2_{\text{SMSE}}\right)^{1/2}$. Where the overall rank was tied, the first model appearing in the table was ranked higher.

$T(\tau)$	$T(m)$	$T(n)$	Kernel	MSLL	SMSE	Rank (MSLL)	Rank (SMSE)	Rank (Total)
$I(\tau)$	$I(m)$	$I(n)$	Exponential	-0.7560	0.2203	8.0	16.0	16.0
		$\log(n)$	Exponential	-0.7875	0.2181	7.0	15.0	15.0
	$\log(m)$	$I(n)$	Exponential	-0.9947	0.1510	3.0	10.0	4.0
		$\log(n)$	Exponential	-0.9846	0.1449	4.0	9.0	3.0
$\log(\tau)$	$I(m)$	$I(n)$	Exponential	-0.8015	0.2132	6.0	14.0	12.0
		$\log(n)$	Exponential	-0.8752	0.1825	5.0	13.0	7.0
	$\log(m)$	$I(n)$	Exponential	-1.0363	0.1442	1.0	8.0	2.0
		$\log(n)$	Exponential	-1.0279	0.1327	2.0	6.0	1.0
$I(\tau)$	$I(m)$	$I(n)$	Matérn 3-2	inf	inf	-	-	-
		$\log(n)$	Matérn 3-2	-0.5648	0.1809	10.0	12.0	13.0
	$\log(m)$	$I(n)$	Matérn 3-2	-0.4921	6.5461	-	-	-
		$\log(n)$	Matérn 3-2	-0.5509	0.1001	11.0	1.0	5.0
$\log(\tau)$	$I(m)$	$I(n)$	Matérn 3-2	17.1108	4.4912	-	-	-
		$\log(n)$	Matérn 3-2	-0.2896	0.1322	13.0	5.0	8.0
	$\log(m)$	$I(n)$	Matérn 3-2	0.8341	6.6805	-	-	-
		$\log(n)$	Matérn 3-2	-0.6497	0.1530	9.0	11.0	9.0
$I(\tau)$	$I(m)$	$I(n)$	Matérn 5-2	0.0998	0.1507	-	-	-
		$\log(n)$	Matérn 5-2	0.2457	0.1419	-	-	-
	$\log(m)$	$I(n)$	Matérn 5-2	-0.2353	0.1103	15.0	2.0	11.0
		$\log(n)$	Matérn 5-2	0.1854	0.0885	-	-	-
$\log(\tau)$	$I(m)$	$I(n)$	Matérn 5-2	0.1737	0.1471	-	-	-
		$\log(n)$	Matérn 5-2	0.1300	0.1468	-	-	-
	$\log(m)$	$I(n)$	Matérn 5-2	-0.2515	0.1228	14.0	4.0	10.0
		$\log(n)$	Matérn 5-2	-0.3690	0.1335	12.0	7.0	6.0
$I(\tau)$	$I(m)$	$I(n)$	Gaussian	0.4745	0.1570	-	-	-
		$\log(n)$	Gaussian	0.3424	0.1426	-	-	-
	$\log(m)$	$I(n)$	Gaussian	-0.0644	0.1120	16.0	3.0	14.0
		$\log(n)$	Gaussian	0.6375	0.0887	-	-	-
$\log(\tau)$	$I(m)$	$I(n)$	Gaussian	0.5639	0.1596	-	-	-
		$\log(n)$	Gaussian	0.9161	0.1642	-	-	-
	$\log(m)$	$I(n)$	Gaussian	0.2483	0.1132	-	-	-
		$\log(n)$	Gaussian	0.1566	0.1366	-	-	-

TABLE 11.5: GAUSSIAN PROCESS MODEL SELECTION METRICS FOR THE RESPONSE SURFACE OF AADH USING HYPERPARAMETER TRIAL DATA SUBSET 4, $\mathscr{D}^4_{100}$. The mean standardised log loss (MSLL) and standardised mean square error (SMSE) where calculated using 10 fold cross validation. Only those models which had both MSLL < 0 and SMSE < 1 were ranked. The total rank is calculated as rank of $\left(R^2_{\text{MSLL}} + R^2_{\text{SMSE}}\right)^{1/2}$. Where the overall rank was tied, the first model appearing in the table was ranked higher.

$T(\tau)$	$T(m)$	$T(n)$	Kernel	MSLL	SMSE	Rank (MSLL)	Rank (SMSE)	Rank (Total)
$I(\tau)$	$I(m)$	$I(n)$	Exponential	-0.4560	0.3075	4.0	16.0	11.0
		$\log(n)$	Exponential	-0.3903	0.3240	8.0	19.0	18.0
	$\log(m)$	$I(n)$	Exponential	-0.2797	0.3010	13.0	15.0	15.0
		$\log(n)$	Exponential	-0.4260	0.2957	5.0	14.0	8.0
$\log(\tau)$	$I(m)$	$I(n)$	Exponential	-0.3979	0.3182	7.0	18.0	12.0
		$\log(n)$	Exponential	-0.3533	0.3126	10.0	17.0	13.0
	$\log(m)$	$I(n)$	Exponential	-0.2961	0.2828	12.0	11.0	10.0
		$\log(n)$	Exponential	-0.1504	0.2821	17.0	10.0	14.0
$I(\tau)$	$I(m)$	$I(n)$	Matérn 3-2	-0.2251	0.2832	16.0	12.0	17.0
		$\log(n)$	Matérn 3-2	-0.4864	0.2769	3.0	9.0	4.0
	$\log(m)$	$I(n)$	Matérn 3-2	-0.3340	0.2395	11.0	2.0	5.0
		$\log(n)$	Matérn 3-2	-0.2605	0.2408	14.0	4.0	7.0
$\log(\tau)$	$I(m)$	$I(n)$	Matérn 3-2	0.6857	0.2878	-	-	-
		$\log(n)$	Matérn 3-2	0.6788	0.2833	-	-	-
	$\log(m)$	$I(n)$	Matérn 3-2	-0.0028	0.2589	19.0	6.0	16.0
		$\log(n)$	Matérn 3-2	2.5955	0.2787	-	-	-
$I(\tau)$	$I(m)$	$I(n)$	Matérn 5-2	-0.0040	0.2865	18.0	13.0	19.0
		$\log(n)$	Matérn 5-2	-0.3897	0.2747	9.0	8.0	6.0
	$\log(m)$	$I(n)$	Matérn 5-2	-0.7173	0.2285	1.0	1.0	1.0
		$\log(n)$	Matérn 5-2	-0.6388	0.2406	2.0	3.0	2.0
$\log(\tau)$	$I(m)$	$I(n)$	Matérn 5-2	0.3359	0.2922	-	-	-
		$\log(n)$	Matérn 5-2	1.9673	0.2817	-	-	-
	$\log(m)$	$I(n)$	Matérn 5-2	-0.2367	0.2533	15.0	5.0	9.0
		$\log(n)$	Matérn 5-2	2.4714	0.2870	-	-	-
$I(\tau)$	$I(m)$	$I(n)$	Gaussian	2.5656	0.2672	-	-	-
		$\log(n)$	Gaussian	-0.4234	0.2669	6.0	7.0	3.0
	$\log(m)$	$I(n)$	Gaussian	1.0865	0.2423	-	-	-
		$\log(n)$	Gaussian	0.2375	0.2507	-	-	-
$\log(\tau)$	$I(m)$	$I(n)$	Gaussian	3.6181	0.2821	-	-	-
		$\log(n)$	Gaussian	1.8240	0.2790	-	-	-
	$\log(m)$	$I(n)$	Gaussian	0.3803	0.2535	-	-	-
		$\log(n)$	Gaussian	6.8263	0.2731	-	-	-

TABLE 11.6: GAUSSIAN PROCESS MODEL SELECTION METRICS FOR THE RESPONSE SURFACE OF AADH USING HYPERPARAMETER TRIAL DATA SUBSET 5, $\mathscr{D}^5_{100}$. The mean standardised log loss (MSLL) and standardised mean square error (SMSE) where calculated using 10 fold cross validation. Only those models which had both MSLL < 0 and SMSE < 1 were ranked. The total rank is calculated as rank of $\left(R^2_{\text{MSLL}} + R^2_{\text{SMSE}}\right)^{1/2}$. Where the overall rank was tied, the first model appearing in the table was ranked higher.

Case	g	d	N_{obs}	g^{s}	Entropy	ICL
Base case	2	620	98,000	1	2.007e+00	1.070e+06
Base case	3	932	98,000	2	5.263e+00	1.009e+06
Base case	4	1,245	98,000	3	8.914e+02	9.375e+05
Base case	5	1,559	98,000	3	2.994e+02	9.036e+05
Base case	6	1,874	98,000	4	9.797e+02	8.715e+05
Base case	7	2,190	98,000	5	2.093e+03	8.491e+05
Base case	8	2,507	98,000	6	2.968e+03	8.342e+05
Base case	9	2,825	98,000	7	3.399e+03	8.312e+05
Base case	10	3,144	98,000	8	3.369e+03	8.083e+05
Base case	11	3,464	98,000	9	4.200e+03	7.966e+05
Base case	12	3,785	98,000	10	5.026e+03	7.847e+05
Base case	13	4,107	98,000	11	4.075e+03	7.936e+05
Base case	14	4,430	98,000	12	5.022e+03	7.891e+05
Base case	15	4,754	98,000	13	5.070e+03	7.854e+05
Base case	16	5,079	98,000	13	5.032e+03	7.870e+05
Base case	17	5,405	98,000	14	5.963e+03	7.801e+05
Base case	18	5,732	98,000	15	6.674e+03	7.785e+05
Base case	19	6,060	98,000	16	6.400e+03	7.786e+05
Base case	20	6,389	98,000	17	9.177e+03	7.806e+05
Sensitivity 1	2	608	80,000	1	2.036e+02	1.040e+06
Sensitivity 1	5	1,529	80,000	1	2.457e+03	9.361e+05
Sensitivity 1	6	1,838	80,000	4	8.101e+03	9.287e+05
Sensitivity 1	8	2,459	80,000	2	9.329e+03	9.233e+05
Sensitivity 2	2	220	98,000	1	2.531e-02	8.224e+05
Sensitivity 2	3	332	98,000	2	7.167e+02	7.139e+05
Sensitivity 2	4	445	98,000	3	1.679e+02	6.509e+05
Sensitivity 2	5	559	98,000	4	1.086e+03	6.165e+05
Sensitivity 2	6	674	98,000	5	1.409e+03	5.951e+05
Sensitivity 2	7	790	98,000	5	1.612e+03	5.874e+05
Sensitivity 2	8	907	98,000	6	1.672e+03	5.776e+05
Sensitivity 2	9	1,025	98,000	6	1.594e+03	5.718e+05
Sensitivity 2	10	1,144	98,000	7	1.873e+03	5.676e+05
Sensitivity 2	11	1,264	98,000	8	2.759e+03	5.637e+05
Sensitivity 2	12	1,385	98,000	8	2.278e+03	5.600e+05
Sensitivity 2	13	1,507	98,000	8	2.382e+03	5.552e+05

Continued on next page

Table 11.6 – continued from previous page

Case	g	d	N_{obs}	g^{s}	Entropy	ICL
Sensitivity 2	14	1,630	98,000	9	2.861e+03	5.493e+05
Sensitivity 2	15	1,754	98,000	10	3.087e+03	5.483e+05
Sensitivity 2	16	1,879	98,000	11	3.710e+03	5.451e+05
Sensitivity 2	17	2,005	98,000	11	8.145e+03	5.369e+05
Sensitivity 2	18	2,132	98,000	11	8.196e+03	5.361e+05
Sensitivity 2	19	2,260	98,000	13	8.662e+03	5.294e+05
Sensitivity 2	20	2,389	98,000	12	8.653e+03	5.299e+05
Sensitivity 3	2	620	98,000	2	8.322e+01	9.751e+05
Sensitivity 3	3	932	98,000	3	1.249e+02	9.290e+05
Sensitivity 3	4	1,245	98,000	4	1.479e+03	9.097e+05
Sensitivity 3	5	1,559	98,000	5	4.028e+03	8.949e+05
Sensitivity 3	6	1,874	98,000	6	3.845e+03	8.834e+05
Sensitivity 3	7	2,190	98,000	7	3.899e+03	8.806e+05
Sensitivity 3	8	2,507	98,000	7	5.300e+03	8.774e+05
Sensitivity 3	9	2,825	98,000	6	4.499e+03	8.720e+05
Sensitivity 3	10	3,144	98,000	6	4.837e+03	8.734e+05
Sensitivity 3	11	3,464	98,000	8	1.090e+04	8.636e+05
Sensitivity 3	12	3,785	98,000	10	9.241e+03	8.664e+05
Sensitivity 3	13	4,107	98,000	12	9.855e+03	8.619e+05
Sensitivity 3	14	4,430	98,000	12	9.923e+03	8.622e+05
Sensitivity 3	15	4,754	98,000	13	1.040e+04	8.613e+05
Sensitivity 3	16	5,079	98,000	14	1.201e+04	8.613e+05
Sensitivity 3	17	5,405	98,000	15	1.240e+04	8.641e+05
Sensitivity 3	18	5,732	98,000	16	1.285e+04	8.612e+05
Sensitivity 3	19	6,060	98,000	17	1.385e+04	8.621e+05
Sensitivity 3	20	6,389	98,000	18	1.401e+04	8.655e+05

TABLE 11.7: INTEGRATED COMPLETE-DATA LIKELIHOOD FOR ALL HMMS. g is the stipulated number of hidden states, g^{s} is the largest strongly connected set of states, d is the number of degrees of freedom, N_{obs} is the number of observations, as given by equation 5.15, ICL and classification entropy are given by equations 5.12 and 5.13 respectively.

i	j	$\hat{R}$	i	j	$\hat{R}$	i	j	$\hat{R}$	i	j	$\hat{R}$
1	1	1.02	8	11	1.01	5	2	1.01	12	8	1.04
1	4	1.00	8	12	1.06	5	3	1.01	12	9	1.00
1	9	1.00	8	13	1.04	5	4	1.05	12	10	1.01
1	10	1.01	9	1	1.00	5	5	1.00	12	11	1.02
1	13	1.00	9	4	1.10	5	6	1.01	12	12	1.03
1	14	1.00	9	5	1.00	5	7	1.00	12	13	1.02
1	15	1.03	9	7	1.07	5	8	1.00	13	1	1.00
2	2	1.03	9	8	1.00	5	9	1.01	13	2	1.00
2	3	1.05	9	9	1.10	5	12	1.00	13	4	1.00
2	5	1.04	9	10	1.00	5	13	1.00	13	5	1.00
2	6	1.00	9	12	1.00	6	2	1.00	13	7	1.01
2	7	1.00	9	13	1.00	6	3	1.00	13	8	1.03
2	8	1.00	10	1	1.01	6	5	1.03	13	9	1.00
2	12	1.00	10	4	1.00	6	6	1.03	13	10	1.00
2	13	1.00	10	8	1.01	7	2	1.00	13	11	1.00
3	2	1.05	10	9	1.00	7	4	1.00	13	12	1.02
3	3	1.00	10	10	1.04	7	5	1.00	13	13	1.02
3	5	1.01	10	12	1.01	7	7	1.03	13	15	1.00
3	6	1.00	10	13	1.00	7	8	1.02	14	1	1.00
4	1	1.00	10	14	1.02	7	9	1.06	14	4	1.00
4	4	1.02	10	15	1.06	7	11	1.00	14	10	1.02
4	5	1.04	11	7	1.00	7	12	1.00	14	14	1.02
4	7	1.00	11	8	1.01	7	13	1.01	14	15	1.02
4	8	1.00	11	11	1.00	8	2	1.00	15	1	1.04
4	9	1.02	11	12	1.02	8	4	1.00	15	4	1.00
4	10	1.02	11	13	1.00	8	5	1.00	15	10	1.09
4	12	1.00	12	2	1.00	8	7	1.03	15	13	1.00
4	13	1.00	12	4	1.00	8	8	1.08	15	14	1.04
4	14	1.00	12	5	1.00	8	9	1.00	15	15	1.07
4	15	1.00	12	7	1.00	8	10	1.01	–	–	–

TABLE 11.8: $\hat{R}$ STATISTICS FOR THE BASE CASE HMM. The model was estimated with four independent MCMC chains with 4000 posterior samples, taken after removing 1000 burn-in samples. Only the non-zero transition matrix elements are shown.

i	j	k_{ij} (μs^{-1})	i	j	k_{ij} (μs^{-1})	i	j	k_{ij} (μs^{-1})
1	2	0.3 (0.0, 1.3)	3	12	0.1 (0.0, 0.2)	7	9	0.0 (0.0, 0.0)
2	1	0.0 (0.0, 0.2)	12	3	0.0 (0.0, 0.1)	9	7	0.0 (0.0, 0.0)
1	3	22.2 (0.5, 83.5)	3	13	3.1 (0.1, 12.6)	7	10	0.5 (0.1, 1.2)
3	1	9.1 (0.8, 28.1)	13	3	0.8 (0.0, 4.2)	10	7	0.4 (0.1, 1.2)
1	4	0.0 (0.0, 0.0)	3	14	0.1 (0.0, 0.4)	7	11	10.5 (5.3, 17.8)
4	1	0.0 (0.0, 0.0)	14	3	0.0 (0.0, 0.1)	11	7	5.6 (2.0, 11.5)
1	5	0.0 (0.0, 0.0)	3	15	0.0 (-0.0, 0.0)	7	12	0.1 (0.0, 0.2)
5	1	0.0 (0.0, 0.0)	15	3	0.0 (-0.0, 0.0)	12	7	0.1 (0.0, 0.2)
1	6	0.0 (0.0, 0.0)	4	5	17.1 (10.7, 25.6)	7	13	0.4 (-0.2, 2.9)
6	1	0.0 (0.0, 0.0)	5	4	12.1 (6.5, 19.8)	13	7	0.3 (-0.2, 2.2)
1	7	0.0 (0.0, 0.0)	4	6	0.1 (0.0, 0.2)	7	14	0.4 (-0.2, 2.4)
7	1	0.0 (0.0, 0.0)	6	4	0.3 (0.0, 0.9)	14	7	0.3 (-0.2, 2.0)
1	8	0.0 (-0.0, 0.1)	4	7	0.0 (-0.3, 1.0)	7	15	0.0 (-0.0, 0.0)
8	1	0.0 (0.0, 0.0)	7	4	0.0 (-0.3, 1.0)	15	7	0.0 (-0.0, 0.0)
1	9	2.6 (-2.1, 22.8)	4	8	0.0 (0.0, 0.0)	8	9	0.0 (-0.0, 0.1)
9	1	1.5 (-1.0, 12.5)	8	4	0.0 (0.0, 0.0)	9	8	0.0 (-0.0, 0.1)
1	10	0.1 (0.0, 0.7)	4	9	0.0 (0.0, 0.0)	8	10	61.1 (22.5, 112.7)
10	1	0.0 (0.0, 0.1)	9	4	0.0 (0.0, 0.0)	10	8	35.6 (8.3, 78.2)
1	11	0.0 (0.0, 0.0)	4	10	0.0 (0.0, 0.0)	8	11	0.2 (0.0, 0.5)
11	1	0.0 (0.0, 0.0)	10	4	0.0 (0.0, 0.0)	11	8	0.1 (0.0, 0.2)
1	12	0.0 (0.0, 0.0)	4	11	0.0 (0.0, 0.0)	8	12	0.5 (-1.0, 1.7)
12	1	0.0 (0.0, 0.0)	11	4	0.0 (0.0, 0.0)	12	8	0.2 (-0.4, 0.8)
1	13	0.1 (0.0, 0.4)	4	12	0.0 (0.0, 0.0)	8	13	0.0 (-0.0, 0.0)
13	1	0.0 (0.0, 0.0)	12	4	0.0 (0.0, 0.0)	13	8	0.0 (-0.0, 0.0)
1	14	0.0 (0.0, 0.0)	4	13	0.0 (0.0, 0.0)	8	14	0.0 (0.0, 0.0)
14	1	0.0 (0.0, 0.0)	13	4	0.0 (0.0, 0.0)	14	8	0.0 (0.0, 0.0)
1	15	0.0 (0.0, 0.0)	4	14	0.0 (0.0, 0.0)	8	15	0.0 (0.0, 0.0)
15	1	0.0 (0.0, 0.0)	14	4	0.0 (0.0, 0.0)	15	8	0.0 (0.0, 0.0)
2	3	5.4 (0.2, 16.8)	4	15	0.0 (0.0, 0.0)	9	10	0.2 (0.0, 0.8)
3	2	12.0 (1.6, 32.2)	15	4	0.0 (0.0, 0.0)	10	9	0.0 (0.0, 0.2)
2	4	0.0 (0.0, 0.0)	5	6	5.3 (1.5, 11.4)	9	11	0.0 (0.0, 0.0)
4	2	0.0 (0.0, 0.0)	6	5	25.8 (4.5, 63.3)	11	9	0.0 (0.0, 0.0)
2	5	0.0 (0.0, 0.0)	5	7	10.9 (5.7, 18.2)	9	12	0.0 (0.0, 0.0)
5	2	0.0 (0.0, 0.0)	7	5	15.0 (6.8, 27.1)	12	9	0.0 (0.0, 0.0)
2	6	0.0 (0.0, 0.0)	5	8	0.1 (0.0, 0.2)	9	13	0.1 (0.0, 0.5)

Continued on next page

Table 11.8 – continued from previous page

i	j	k_{ij} (μs^{-1})	i	j	k_{ij} (μs^{-1})	i	j	k_{ij} (μs^{-1})
6	2	0.0 (0.0, 0.0)	8	5	0.2 (0.0, 0.5)	13	9	0.0 (0.0, 0.1)
2	7	0.0 (0.0, 0.0)	5	9	0.0 (0.0, 0.0)	9	14	0.0 (0.0, 0.0)
7	2	0.0 (0.0, 0.0)	9	5	0.0 (0.0, 0.0)	14	9	0.0 (0.0, 0.0)
2	8	0.0 (0.0, 0.0)	5	10	0.0 (0.0, 0.0)	9	15	0.0 (0.0, 0.0)
8	2	0.0 (0.0, 0.0)	10	5	0.0 (0.0, 0.0)	15	9	0.0 (0.0, 0.0)
2	9	0.1 (0.0, 0.5)	5	11	0.1 (0.1, 0.3)	10	11	1.0 (-0.1, 4.5)
9	2	0.4 (0.0, 1.3)	11	5	0.1 (0.0, 0.2)	11	10	0.7 (-0.1, 2.9)
2	10	0.0 (0.0, 0.2)	5	12	0.0 (0.0, 0.0)	10	12	10.2 (2.8, 22.5)
10	2	0.0 (0.0, 0.1)	12	5	0.0 (0.0, 0.0)	12	10	7.1 (1.8, 16.7)
2	11	0.0 (0.0, 0.0)	5	13	0.0 (-0.0, 0.0)	10	13	0.0 (-0.0, 0.0)
11	2	0.0 (0.0, 0.0)	13	5	0.0 (-0.0, 0.0)	13	10	0.0 (-0.0, 0.0)
2	12	0.0 (0.0, 0.0)	5	14	0.0 (-0.0, 0.0)	10	14	0.1 (0.0, 0.2)
12	2	0.0 (0.0, 0.0)	14	5	0.0 (-0.0, 0.0)	14	10	0.0 (0.0, 0.2)
2	13	0.0 (0.0, 0.1)	5	15	0.0 (0.0, 0.0)	10	15	0.0 (0.0, 0.0)
13	2	0.0 (0.0, 0.1)	15	5	0.0 (0.0, 0.0)	15	10	0.0 (0.0, 0.0)
2	14	0.0 (0.0, 0.0)	6	7	7.2 (0.1, 23.1)	11	12	10.8 (4.8, 19.5)
14	2	0.0 (0.0, 0.0)	7	6	2.0 (0.0, 6.2)	12	11	11.1 (3.5, 23.6)
2	15	0.0 (0.0, 0.0)	6	8	0.1 (-0.0, 0.2)	11	13	8.2 (3.0, 16.1)
15	2	0.0 (0.0, 0.0)	8	6	0.0 (-0.0, 0.1)	13	11	9.4 (2.3, 21.5)
3	4	0.0 (0.0, 0.0)	6	9	0.0 (0.0, 0.0)	11	14	7.4 (1.9, 16.4)
4	3	0.0 (0.0, 0.0)	9	6	0.0 (0.0, 0.0)	14	11	10.5 (2.1, 25.6)
3	5	0.0 (0.0, 0.0)	6	10	0.0 (-0.0, 0.0)	11	15	0.1 (0.0, 0.2)
5	3	0.0 (0.0, 0.0)	10	6	0.0 (0.0, 0.0)	15	11	0.1 (0.0, 0.2)
3	6	0.0 (0.0, 0.0)	6	11	2.5 (-0.1, 10.9)	12	13	0.2 (0.1, 0.6)
6	3	0.0 (0.0, 0.0)	11	6	0.4 (-0.0, 1.9)	13	12	0.2 (0.1, 0.7)
3	7	0.0 (-0.0, 0.1)	6	12	0.0 (-0.0, 0.1)	12	14	4.3 (0.4, 12.9)
7	3	0.0 (0.0, 0.0)	12	6	0.0 (-0.0, 0.0)	14	12	5.8 (0.6, 16.2)
3	8	0.2 (-0.9, 4.4)	6	13	0.0 (-0.0, 0.1)	12	15	0.0 (0.0, 0.1)
8	3	0.1 (-0.6, 2.4)	13	6	0.0 (-0.0, 0.0)	15	12	0.0 (0.0, 0.1)
3	9	21.8 (6.3, 48.2)	6	14	0.0 (-0.0, 0.1)	13	14	24.8 (7.0, 54.6)
9	3	31.5 (6.5, 72.6)	14	6	0.0 (-0.0, 0.0)	14	13	30.3 (10.6, 64.7)
3	10	5.9 (0.3, 19.7)	6	15	0.0 (0.0, 0.0)	13	15	2.3 (-0.4, 9.4)
10	3	2.0 (0.0, 8.5)	15	6	0.0 (0.0, 0.0)	15	13	1.5 (-0.3, 8.6)
3	11	0.0 (0.0, 0.1)	7	8	7.9 (3.3, 14.4)	14	15	6.4 (0.9, 15.5)

Continued on next page

Table 11.8 – continued from previous page

i	j	k_{ij} (μs^{-1})	i	j	k_{ij} (μs^{-1})	i	j	k_{ij} (μs^{-1})
11	3	0.0 (0.0, 0.0)	8	7	11.6 (2.8, 27.8)	15	14	3.8 (0.0, 15.4)

TABLE 11.9: RATE MATRIX FOR BASE CASE HMM. The model was estimated with four independent MCMC chains with 4000 posterior samples, taken after removing 1000 burn-in samples. Values reported are the mean and 95 % credible intervals.

i	$\widetilde{\pi}_i$ (%)
1	3.2 (0.0, 44.3)
2	10.3 (0.1, 73.4)
3	2.5 (0.1, 12.9)
4	4.5 (0.4, 13.3)
5	6.3 (0.5, 16.7)
6	1.6 (0.1, 5.8)
7	4.5 (0.4, 10.6)
8	3.5 (0.3, 10.3)
9	2.3 (0.0, 14.8)
10	6.4 (0.6, 18.2)
11	8.4 (0.8, 17.8)
12	9.0 (0.9, 22.8)
13	7.9 (0.9, 19.9)
14	6.3 (0.7, 15.3)
15	23.4 (1.6, 91.3)

TABLE 11.10: STATIONARY DISTRIBUTION OF BASE CASE HMM. The model was estimated with four independent MCMC chains with 4000 posterior samples, taken after removing 1000 burn-in samples. Values reported are the mean and 95 % credible intervals.

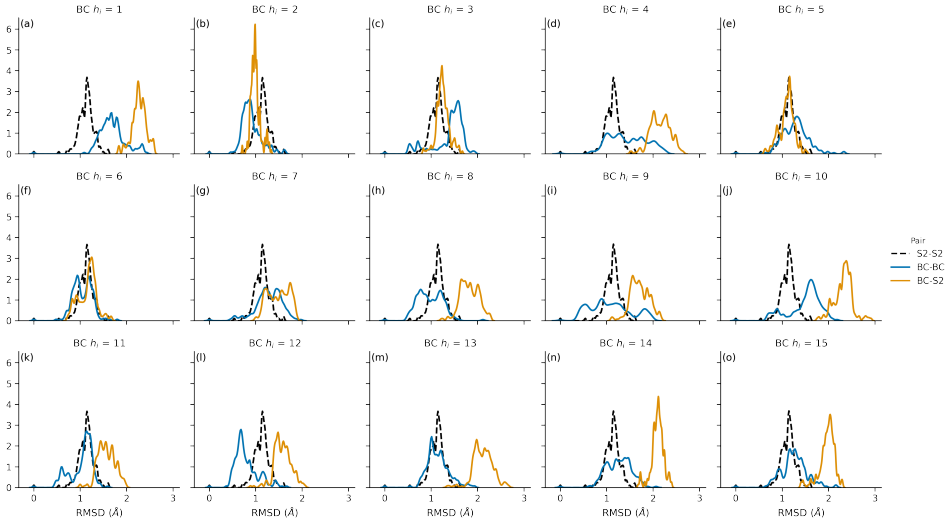


FIGURE 11.7: COMPARISON OF SENSITIVITY 2 HIDDEN STATE h_5 AND THE BASE CASE HIDDEN STATES. Each distribution is a comparison of two hidden states, if the distributions all overlap then the hidden states described the same ensemble of configurations. Each panel compares sensitivity 2 h_5 with a different base case hidden state, h_i. The observations which make up the distributions are heavy atom RMSD between unique pairs of configurations, one from each hidden state. Black dashed line compares sensitivity 2 h_5 with itself (S2-S2), blue solid line compares each base case hidden state with itself (BC-BC). Yellow solid lines compare sensitivity 2 h_5 with each base case hidden state (BC-S2).

i	j	$\hat{R}$	i	j	$\hat{R}$	i	j	$\hat{R}$
1	1	1.01	4	13	1.00	9	6	1.06
1	2	1.03	5	1	1.01	9	7	1.00
1	3	1.00	5	2	1.01	9	8	1.01
1	4	1.00	5	3	1.02	9	9	1.01
1	5	1.00	5	5	1.13	9	10	1.01
1	6	1.00	5	12	1.14	9	11	1.02
1	7	1.00	6	1	1.01	9	13	1.01
1	8	1.00	6	2	1.00	10	1	1.00
1	9	1.00	6	3	1.01	10	2	1.00
1	10	1.00	6	4	1.02	10	4	1.00
1	11	1.00	6	6	1.07	10	6	1.17
1	12	1.00	6	7	1.03	10	7	1.17
2	1	1.02	6	8	1.01	10	8	1.06
2	2	1.12	6	9	1.07	10	9	1.01
2	3	1.10	6	10	1.04	10	10	1.01
2	4	1.00	6	11	1.19	10	11	1.16
2	5	1.00	6	13	1.01	10	13	1.00
2	6	1.04	7	1	1.00	11	1	1.00
2	7	1.11	7	2	1.10	11	2	1.00
2	8	1.00	7	3	1.00	11	3	1.00
2	9	1.00	7	4	1.00	11	4	1.00
2	10	1.00	7	6	1.21	11	6	1.28
2	11	1.00	7	7	1.04	11	7	1.30
2	12	1.04	7	8	1.00	11	8	1.01
3	1	1.00	7	9	1.00	11	9	1.02
3	2	1.04	7	10	1.16	11	10	1.15
3	3	1.05	7	11	1.28	11	11	1.03
3	5	1.08	7	13	1.00	11	13	1.00
3	6	1.00	8	1	1.00	12	1	1.01
3	7	1.00	8	2	1.00	12	2	1.08
3	11	1.00	8	4	1.00	12	3	1.18
3	12	1.11	8	6	1.00	12	5	1.13
4	1	1.00	8	7	1.00	12	12	1.17
4	2	1.00	8	8	1.03	13	4	1.00
4	4	1.00	8	9	1.03	13	6	1.00
4	6	1.02	8	10	1.05	13	7	1.00
4	7	1.00	8	11	1.02	13	8	1.00
4	8	1.00	8	13	1.00	13	9	1.01
4	9	1.00	9	1	1.00	13	10	1.00
4	10	1.00	9	2	1.00	13	11	1.00
4	11	1.00	9	4	1.00	13	13	1.01

TABLE 11.11: $\hat{R}$ STATISTICS FOR SENSITIVITY 2 HMM. The model was estimated with four independent MCMC chains with 4000 posterior samples, taken after removing 1000 burn-in samples. Only the non-zero transition matrix elements are shown.

i	j	k_{ij} (μs^{-1})	i	j	k_{ij} (μs^{-1})	i	j	k_{ij} (μs^{-1})
1	2	40.79 (6.11, 109.70)	7	5	9.15 (1.83, 20.66)	8	10	2.79 (-0.29, 19.49)
2	1	20.65 (3.55, 53.95)	6	8	5.84 (0.12, 23.07)	10	8	0.61 (-0.06, 4.84)
1	4	2.89 (-0.17, 13.81)	8	6	1.86 (0.01, 10.34)	9	10	18.36 (6.40, 35.54)
4	1	6.79 (-0.44, 25.34)	6	10	0.32 (0.03, 1.10)	10	9	12.50 (4.16, 24.47)
2	3	6.44 (0.18, 21.01)	10	6	0.02 (0.00, 0.08)	9	11	2.04 (-0.59, 9.87)
3	2	28.28 (5.97, 65.30)	6	12	12.49 (1.21, 38.17)	11	9	1.72 (-0.49, 8.72)
2	4	1.46 (-1.09, 7.85)	12	6	0.58 (0.01, 2.49)	9	12	1.57 (-0.99, 10.84)
4	2	6.60 (-4.25, 27.34)	7	8	5.07 (0.70, 14.02)	12	9	0.83 (-0.53, 5.58)
3	4	83.47 (16.22, 169.61)	8	7	9.85 (0.98, 30.24)	9	13	0.18 (-0.87, 0.72)
4	3	77.98 (25.08, 149.90)	7	9	18.49 (6.72, 34.04)	13	9	0.24 (-1.15, 0.96)
3	6	11.26 (0.77, 37.39)	9	7	9.53 (2.32, 21.49)	10	12	20.12 (3.23, 49.00)
6	3	10.95 (0.21, 43.49)	7	10	0.14 (-3.21, 1.05)	12	10	16.32 (2.69, 37.19)
3	10	10.27 (0.69, 31.51)	10	7	0.03 (-1.26, 0.41)	10	13	16.03 (6.09, 31.32)
10	3	0.61 (0.00, 3.35)	7	11	26.59 (12.77, 45.31)	13	10	30.40 (8.24, 61.50)
3	13	0.85 (-0.59, 10.02)	11	7	11.17 (2.90, 25.71)	11	12	8.24 (1.37, 21.22)
13	3	0.09 (-0.11, 1.08)	7	12	9.14 (2.53, 19.37)	12	11	5.53 (1.14, 14.25)
4	13	0.06 (-0.07, 0.72)	12	7	2.59 (0.50, 6.34)	12	13	2.31 (-0.27, 7.07)
13	4	0.01 (-0.01, 0.09)	8	9	5.44 (-0.09, 26.09)	13	12	5.46 (-0.55, 17.99)
5	7	1.20 (0.32, 2.58)	9	8	1.73 (-0.02, 8.58)	–	–	–

TABLE 11.12: RATE MATRIX FOR SENSITIVITY 2 HMM. The model was estimated with four independent MCMC chains with 4000 posterior samples, taken after removing 1000 burn-in samples. Values reported are the mean and 95 % credible intervals.

i	$\widetilde{\pi}_i$ (%)
1	3.43 (0.02, 29.98)
2	5.85 (0.04, 43.58)
3	0.73 (0.01, 4.14)
4	0.79 (0.01, 5.19)
5	30.12 (5.27, 68.33)
6	0.84 (0.02, 5.07)
7	3.76 (1.16, 7.50)
8	2.99 (0.31, 14.84)
9	8.07 (2.07, 18.01)
10	11.82 (2.96, 24.86)
11	10.36 (2.82, 25.81)
12	14.33 (4.19, 29.52)
13	6.92 (1.46, 20.01)

TABLE 11.13: STATIONARY DISTRIBUTION OF SENSITIVITY 2 HMM. The model was estimated with four independent MCMC chains with 4000 posterior samples, taken after removing 1000 burn-in samples. Values reported are the mean and 95 % credible intervals.

BIBLIOGRAPHY

[1] R. Zwanzig, "Diffusion in a rough potential,"
Proceedings of the National Academy of Sciences, vol. 85, no. 7, pp. 2029–2030, Apr. 1988.
DOI: 10.1073/pnas.85.7.2029.

[2] S. Hammes-Schiffer and S. J. Benkovic, "Relating Protein Motion to Catalysis,"
Annual Review of Biochemistry, vol. 75, no. 1, pp. 519–541, 2006.
DOI: 10.1146/annurev.biochem.75.103004.142800.

[3] M. Karplus and J. Kuriyan, "Molecular dynamics and protein function,"
Proceedings of the National Academy of Sciences, vol. 102, no. 19, pp. 6679–6685, May 2005.
DOI: 10.1073/pnas.0408930102.

[4] G. G. Hammes, "Multiple Conformational Changes in Enzyme Catalysis,"
Biochemistry, vol. 41, no. 26, pp. 8221–8228, Jul. 2002.
DOI: 10.1021/bi0260839.

[5] M. Roca, B. Messer, D. Hilvert, and A. Warshel, "On the relationship between folding and chemical landscapes in enzyme catalysis,"
Proceedings of the National Academy of Sciences, vol. 105, no. 37, pp. 13 877–13 882, 2008.
DOI: 10.1073/pnas.0803405105.

[6] Q. Liao, Y. Kulkarni, U. Sengupta, D. Petrović, A. J. Mulholland, M. W. van der Kamp, B. Strodel, and S. C. L. Kamerlin, "Loop motion in triosephosphate isomerase is not a simple open and shut case,"
Journal of the American Chemical Society, vol. 140, no. 46, pp. 15 889–15 903, 2018.
DOI: 10.1021/jacs.8b09378.

[7] J. P. Richard, "A Paradigm for Enzyme-Catalyzed Proton Transfer at Carbon: Triosephosphate Isomerase,"
Biochemistry, vol. 51, no. 13, pp. 2652–2661, Apr. 2012.
DOI: 10.1021/bi300195b.

[8] F. K. Brown and P. A. Kollman, "Molecular dynamics simulations of "loop closing" in the enzyme triose phosphate isomerase,"
Journal of Molecular Biology, vol. 198, no. 3, pp. 533–546, Dec. 1987.
DOI: 10.1016/0022-2836(87)90298-1.

[9] D. Joseph, G. A. Petsko, and M. Karplus, “Anatomy of a conformational change: Hinged "lid" motion of the triosephosphate isomerase loop,”
Science, vol. 249, no. 4975, pp. 1425–1428, Sep. 1990.
DOI: 10.1126/science.2402636.

[10] J. R. Schnell, H. J. Dyson, and P. E. Wright, “Structure, Dynamics, and Catalytic Function of Dihydrofolate Reductase,”
Annual Review of Biophysics and Biomolecular Structure, vol. 33, no. 1, pp. 119–140, 2004.
DOI: 10.1146/annurev.biophys.33.110502.133613.

[11] C. A. Fierke, K. A. Johnson, and S. J. Benkovic, “Construction and evaluation of the kinetic scheme associated with dihydrofolate reductase from Escherichia coli,”
Biochemistry, vol. 26, no. 13, pp. 4085–4092, Jun. 1987.
DOI: 10.1021/bi00387a052.

[12] M. R. Sawaya and J. Kraut, “Loop and Subdomain Movements in the Mechanism of Escherichia coli Dihydrofolate Reductase: Crystallographic Evidence,”
Biochemistry, vol. 36, no. 3, pp. 586–603, Jan. 1997.
DOI: 10.1021/bi962337c.

[13] L. Li, P. E. Wright, S. J. Benkovic, and C. J. Falzone, “Functional role of a mobile loop of Escherichia coli dihydrofolate reductase in transition-state stabilization,”
Biochemistry, vol. 31, no. 34, pp. 7826–7833, Sep. 1992.
DOI: 10.1021/bi00149a012.

[14] D. G. Truhlar and A. Kohen, “Convex Arrhenius plots and their interpretation,”
Proceedings of the National Academy of Sciences, vol. 98, no. 3, pp. 848–851, Jan. 2001.
DOI: 10.1073/pnas.98.3.848.

[15] R. M. Daniel and M. J. Danson, “A new understanding of how temperature affects the catalytic activity of enzymes,”
Trends in Biochemical Sciences, vol. 35, no. 10, pp. 584–591, Oct. 2010.
DOI: 10.1016/j.tibs.2010.05.001.

[16] J. Åqvist, J. Sočan, and M. Purg, “Hidden Conformational States and Strange Temperature Optima in Enzyme Catalysis,”
Biochemistry, vol. 59, no. 40, pp. 3844–3855, Oct. 2020.
DOI: 10.1021/acs.biochem.0c00705.

[17] V. L. Arcus and A. J. Mulholland, “Temperature, Dynamics, and Enzyme-Catalyzed Reaction Rates,”
Annual Review of Biophysics, vol. 49, no. 1, pp. 163–180, 2020.
DOI: 10.1146/annurev-biophys-121219-081520.

[18] M. A. Maria-Solano, E. Serrano-Hervás, A. Romero-Rivera, J. Iglesias-Fernández, and S. Osuna, "Role of conformational dynamics in the evolution of novel enzyme function," *Chemical Communications*, vol. 54, no. 50, pp. 6622–6634, Jun. 2018. DOI: 10.1039/C8CC02426J.

[19] E. Campbell, M. Kaltenbach, G. J. Correy, P. D. Carr, B. T. Porebski, E. K. Livingstone, L. Afriat-Jurnou, A. M. Buckle, M. Weik, F. Hollfelder, N. Tokuriki, and C. J. Jackson, "The role of protein dynamics in the evolution of new enzyme function," *Nature Chemical Biology*, vol. 12, no. 11, pp. 944–950, Nov. 2016. DOI: 10.1038/nchembio.2175.

[20] A. Romero-Rivera, M. Garcia-Borràs, and S. Osuna, "Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity," *ACS Catalysis*, vol. 7, no. 12, pp. 8524–8532, Dec. 2017. DOI: 10.1021/acscatal.7b02954.

[21] D. M. Rosenbaum, S. G. F. Rasmussen, and B. K. Kobilka, "The structure and function of G-protein-coupled receptors," *Nature*, vol. 459, no. 7245, pp. 356–363, May 2009. DOI: 10.1038/nature08144.

[22] S. Bockenhauer, A. Fürstenberg, X. J. Yao, B. K. Kobilka, and W. E. Moerner, "Conformational Dynamics of Single G Protein-Coupled Receptors in Solution," *The Journal of Physical Chemistry B*, vol. 115, no. 45, pp. 13 328–13 338, Nov. 2011. DOI: 10.1021/jp204843r.

[23] J. Liu and R. Nussinov, "Allostery: An Overview of Its History, Concepts, Methods, and Applications," *PLoS Computational Biology*, vol. 12, no. 6, e1004966, Jun. 2016. DOI: 10.1371/journal.pcbi.1004966.

[24] W. N. Lipscomb and E. R. Kantrowitz, "Structure and Mechanisms of Escherichia coli Aspartate Transcarbamoylase," *Accounts of Chemical Research*, vol. 45, no. 3, pp. 444–453, Mar. 2012. DOI: 10.1021/ar200166p.

[25] J.-P. Changeux and A. Christopoulos, "Allosteric Modulation as a Unifying Mechanism for Receptor Function and Regulation," *Cell*, vol. 166, no. 5, pp. 1084–1102, Aug. 2016. DOI: 10.1016/j.cell.2016.08.015.

[26] A. van der Vaart, "Coupled binding–bending–folding: The complex conformational dynamics of protein-DNA binding studied by atomistic molecular dynamics simulations," *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1850, no. 5, pp. 1091–1098, May 2015.

DOI: 10.1016/j.bbagen.2014.08.009.

[27] D. J. Huggins, P. C. Biggin, M. A. Dämgen, J. W. Essex, S. A. Harris, R. H. Henchman, S. Khalid, A. Kuzmanic, C. A. Laughton, J. Michel, A. J. Mulholland, E. Rosta, M. S. P. Sansom, and M. W. van der Kamp, "Biomolecular simulations: From dynamics and mechanisms to computational assays of biological activity,"
WIREs Computational Molecular Science, vol. 9, no. 3, e1393, 2019.
DOI: 10.1002/wcms.1393.

[28] M. A. Rohrdanz, W. Zheng, and C. Clementi, "Discovering Mountain Passes via Torchlight: Methods for the Definition of Reaction Coordinates and Pathways in Complex Macromolecular Reactions,"
Annual Review of Physical Chemistry, vol. 64, no. 1, pp. 295–316, Apr. 2013.
DOI: 10.1146/annurev-physchem-040412-110006.

[29] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, "TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark,"
Annual Review of Physical Chemistry, vol. 53, no. 1, pp. 291–318, 2002.
DOI: 10.1146/annurev.physchem.53.082301.113146.

[30] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, "Transition path sampling and the calculation of rate constants,"
The Journal of Chemical Physics, vol. 108, no. 5, pp. 1964–1977, Feb. 1998.
DOI: 10.1063/1.475562.

[31] C. Dellago, P. G. Bolhuis, and P. L. Geissler,
"Transition Path Sampling,"
In *Advances in Chemical Physics*,
John Wiley & Sons, Ltd, 2002,
Pp. 1–78.

[32] T. S. van Erp, D. Moroni, and P. G. Bolhuis, "A novel path sampling method for the calculation of rate constants,"
The Journal of Chemical Physics, vol. 118, no. 17, pp. 7762–7774, Apr. 2003.
DOI: 10.1063/1.1562614.

[33] R. J. Allen, P. B. Warren, and P. R. ten Wolde, "Sampling Rare Switching Events in Biochemical Networks,"
Physical Review Letters, vol. 94, no. 1, p. 018 104, Jan. 2005.
DOI: 10.1103/PhysRevLett.94.018104.

[34] J. Juraszek and P. G. Bolhuis, "Sampling the multiple folding mechanisms of Trp-cage in explicit solvent,"
Proceedings of the National Academy of Sciences, vol. 103, no. 43, pp. 15 859–15 864, Oct. 2006.
DOI: 10.1073/pnas.0606692103.

[35] J. Juraszek and P. G. Bolhuis, "Rate Constant and Reaction Coordinate of Trp-Cage Folding in Explicit Water,"
Biophysical Journal, vol. 95, no. 9, pp. 4246–4257, Nov. 2008.
DOI: 10.1529/biophysj.108.136267.

[36] C. Velez-Vega, E. E. Borrero, and F. A. Escobedo, "Kinetics and mechanism of the unfolding native-to-loop transition of Trp-cage in explicit solvent via optimized forward flux sampling simulations,"
The Journal of Chemical Physics, vol. 133, no. 10, p. 105 103, Sep. 2010.
DOI: 10.1063/1.3474803.

[37] J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen, "Designing a 20-residue protein,"
Nature Structural Biology, vol. 9, no. 6, pp. 425–430, Jun. 2002.
DOI: 10.1038/nsb798.

[38] Weinan, E. and Vanden-Eijnden, E., "Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events,"
Annual Review of Physical Chemistry, vol. 61, pp. 391–420, 2010.
DOI: 10.1146/annurev.physchem.040808.090412.

[39] H. J?nsson, G. Mills, and K. W. Jacobsen,
"Nudged elastic band method for finding minimum energy paths of transitions,"
In *Classical and Quantum Dynamics in Condensed Phase Simulations*,
WORLD SCIENTIFIC, Jun. 1998,
Pp. 385–404.
DOI: 10.1142/9789812839664_0016.

[40] R. M. Elder and A. Jayaraman, "Sequence-Specific Recognition of Cancer Drug-DNA Adducts by HMGB1a Repair Protein,"
Biophysical Journal, vol. 102, no. 10, pp. 2331–2338, May 2012.
DOI: 10.1016/j.bpj.2012.04.013.

[41] A. K. Faradjian and R. Elber, "Computing time scales from reaction coordinates by milestoning,"
The Journal of Chemical Physics, vol. 120, no. 23, pp. 10 880–10 889, Jun. 2004.
DOI: 10.1063/1.1738640.

[42] P. Májek and R. Elber, "Milestoning without a Reaction Coordinate,"
Journal of Chemical Theory and Computation, vol. 6, no. 6, pp. 1805–1817, Jun. 2010.
DOI: 10.1021/ct100114j.

[43] S. Kirmizialtin and R. Elber, "Revisiting and Computing Reaction Coordinates with Directional Milestoning,"
The Journal of Physical Chemistry A, vol. 115, no. 23, pp. 6137–6148, Jun. 2011.
DOI: 10.1021/jp111093c.

[44] S. Kirmizialtin, V. Nguyen, K. A. Johnson, and R. Elber, "How Conformational Dynamics of DNA Polymerase Select Correct Substrates: Experiments and Simulations,"
Structure, vol. 20, no. 4, pp. 618–627, Apr. 2012.
DOI: 10.1016/j.str.2012.02.018.

[45] A. E. Cardenas, G. S. Jas, K. Y. DeLeon, W. A. Hegefeld, K. Kuczera, and R. Elber, "Unassisted Transport of *N* -Acetyl- L -tryptophanamide through Membrane: Experiment and Simulation of Kinetics,"
The Journal of Physical Chemistry B, vol. 116, no. 9, pp. 2739–2750, Mar. 2012.
DOI: 10.1021/jp2102447.

[46] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space,"
The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 2, no. 11, pp. 559–572, 1901.

[47] S. B. Nolde, A. S. Arseniev, V. Y. Orekhov, and M. Billeter, "Essential domain motions in barnase revealed by MD simulations,"
Proteins: Structure, Function, and Bioinformatics, vol. 46, no. 3, pp. 250–258, 2002.
DOI: 10.1002/prot.10030.

[48] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, "Essential dynamics of proteins,"
Proteins: Structure, Function, and Bioinformatics, vol. 17, no. 4, pp. 412–425, 1993.
DOI: 10.1002/prot.340170408.

[49] D. Antoniou and S. D. Schwartz, "Toward Identification of the Reaction Coordinate Directly from the Transition State Ensemble Using the Kernel PCA Method,"
The Journal of Physical Chemistry B, vol. 115, no. 10, pp. 2465–2469, Mar. 2011.
DOI: 10.1021/jp111682x.

[50] S. L. Quaytman and S. D. Schwartz, "Reaction coordinate of an enzymatic reaction revealed by transition path sampling,"
Proceedings of the National Academy of Sciences, vol. 104, no. 30, pp. 12 253–12 258, Jul. 2007.
DOI: 10.1073/pnas.0704304104.

[51] I. Borg and P. Groenen,
Modern Multidimensional Scaling: Theory and Applications.
Springer, 1997.

[52] J. M. Troyer and F. E. Cohen, "Protein conformational landscapes: Energy minimization and clustering of a long molecular dynamics trajectory,"
Proteins: Structure, Function, and Bioinformatics, vol. 23, no. 1, pp. 97–110, 1995.
DOI: 10.1002/prot.340230111.

[53] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction,"

Science, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.
DOI: 10.1126/science.290.5500.2319.

[54] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, "Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction,"
Proceedings of the National Academy of Sciences, vol. 103, no. 26, pp. 9885–9890, Jun. 2006.
DOI: 10.1073/pnas.0603553103.

[55] M. Ceriotti, G. A. Tribello, and M. Parrinello, "Simplifying the representation of complex free-energy landscapes using sketch-map,"
Vol. 108, no. 32, pp. 13 023–13 028, 2011.
DOI: 10.1073/pnas.1108486108.

[56] A. Ardevol, G. A. Tribello, M. Ceriotti, and M. Parrinello, "Probing the unfolded configurations of a β-hairpin using sketch-map,"
Journal of Chemical Theory and Computation, vol. 11, no. 3, pp. 1086–1093, 2015.
DOI: 10.1021/ct500950z.

[57] A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, and P. G. Debenedetti, "Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach,"
Chemical Physics Letters, vol. 509, no. 1, pp. 1–11, Jun. 2011.
DOI: 10.1016/j.cplett.2011.04.066.

[58] S. B. Kim, C. J. Dsilva, I. G. Kevrekidis, and P. G. Debenedetti, "Systematic characterization of protein folding pathways using diffusion maps: Application to Trp-cage miniprotein,"
The Journal of Chemical Physics, vol. 142, no. 8, p. 085 101, Feb. 2015.
DOI: 10.1063/1.4913322.

[59] W. Zheng, B. Qi, M. A. Rohrdanz, A. Caflisch, A. R. Dinner, and C. Clementi, "Delineation of Folding Pathways of a β-Sheet Miniprotein,"
The Journal of Physical Chemistry B, vol. 115, no. 44, pp. 13 065–13 074, Nov. 2011.
DOI: 10.1021/jp2076935.

[60] B. E. Husic and V. S. Pande, "Markov State Models: From an Art to a Science,"
Journal of the American Chemical Society, vol. 140, no. 7, pp. 2386–2396, Feb. 2018.
DOI: 10.1021/jacs.7b12191.

[61] V. S. Pande, K. Beauchamp, and G. R. Bowman, "Everything you wanted to know about Markov State Models but were afraid to ask,"
Methods, vol. 52, no. 1, pp. 99–105, Sep. 2010.
DOI: 10.1016/j.ymeth.2010.06.002.

[62] R. Zwanzig, "From classical dynamics to continuous time random walks,"
Journal of Statistical Physics, vol. 30, no. 2, pp. 255–262, Feb. 1983.
DOI: 10.1007/BF01012300.

[63] F. Noé and S. Fischer, "Transition networks for modeling the kinetics of conformational change in macromolecules,"
Current Opinion in Structural Biology, vol. 18, no. 2, pp. 154–162, 2008.
DOI: 10.1016/j.sbi.2008.01.008.

[64] J. H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schtte, and F. Noé, "Markov models of molecular kinetics: Generation and validation,"
Journal of Chemical Physics, vol. 134, no. 17, 2011.
DOI: 10.1063/1.3565032.

[65] N. Singhal, C. D. Snow, and V. S. Pande, "Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin,"
The Journal of Chemical Physics, vol. 121, no. 1, p. 415, 2004.
DOI: 10.1063/1.1738647.

[66] W. C. Swope, J. W. Pitera, and F. Suits, "Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory,"
The Journal of Physical Chemistry B, vol. 108, no. 21, pp. 6571–6581, May 2004.
DOI: 10.1021/jp037421y.

[67] M. Schor, A. S. J. S. Mey, and C. E. MacPhee, "Analytical methods for structural ensembles and dynamics of intrinsically disordered proteins,"
Biophysical Reviews, vol. 8, no. 4, pp. 429–439, Dec. 2016.
DOI: 10.1007/s12551-016-0234-6.

[68] N. Ahalawat and J. Mondal, "Mapping the Substrate Recognition Pathway in Cytochrome P450,"
Journal of the American Chemical Society, vol. 140, no. 50, pp. 17 743–17 752, Dec. 2018.
DOI: 10.1021/jacs.8b10840.

[69] S. C. Dodani, G. Kiss, J. K. B. Cahn, Y. Su, V. S. Pande, and F. H. Arnold, "Discovery of a regioselectivity switch in nitrating P450s guided by MD simulations and Markov models,"
Nature chemistry, vol. 8, no. 5, pp. 419–425, May 2016.
DOI: 10.1038/nchem.2474.

[70] S. Chen, R. P. Wiewiora, F. Meng, N. Babault, A. Ma, W. Yu, K. Qian, H. Hu, H. Zou, J. Wang, S. Fan, G. Blum, F. Pittella-Silva, K. A. Beauchamp, W. Tempel, H. Jiang, K. Chen, R. J. Skene, Y. G. Zheng, P. J. Brown, J. Jin, C. Luo, J. D. Chodera, and M. Luo, "The dynamic conformational landscape of the protein methyltransferase SETD8,"
eLife, vol. 8, S. J. Fleishman, J. Kuriyan, G. R. Bowman, and E. Lindahl, Eds., e45403, May 2019.
DOI: 10.7554/eLife.45403.

[71] P. Wapeesittipan, A. S. J. S. Mey, M. D. Walkinshaw, and J. Michel, "Allosteric effects in cyclophilin mutants may be explained by changes in nano-microsecond time scale motions,"
Communications Chemistry, vol. 2, no. 1, pp. 1–9, Mar. 2019.

DOI: 10.1038/s42004-019-0136-1.

[72] U. Sengupta, M. Carballo-Pacheco, and B. Strodel, "Automated Markov state models for molecular dynamics simulations of aggregation and self-assembly,"
The Journal of Chemical Physics, vol. 150, no. 11, p. 115 101, Mar. 2019.
DOI: 10.1063/1.5083915.

[73] C. T. Leahy, R. D. Murphy, G. Hummer, E. Rosta, and N.-V. Buchete, "Coarse Master Equations for Binding Kinetics of Amyloid Peptide Dimers,"
The Journal of Physical Chemistry Letters, vol. 7, no. 14, pp. 2676–2682, Jul. 2016.
DOI: 10.1021/acs.jpclett.6b00518.

[74] R. E. Amaro, J. Baudry, J. Chodera, Ö. Demir, J. A. McCammon, Y. Miao, and J. C. Smith, "Ensemble Docking in Drug Discovery,"
Biophysical Journal, vol. 114, no. 10, pp. 2271–2278, May 2018.
DOI: 10.1016/j.bpj.2018.02.038.

[75] F. L. Gervasio, V. Spiwok, and R. Mannhold,
Biomolecular Simulations in Structure-Based Drug Discovery.
John Wiley & Sons, Apr. 2019.
DOI: 10.1002/9783527806836.

[76] D. R. Glowacki, C.-H. Liang, C. Morley, M. J. Pilling, and S. H. Robertson, "MESMER: An Open-Source Master Equation Solver for Multi-Energy Well Reactions,"
The Journal of Physical Chemistry A, vol. 116, no. 38, pp. 9545–9560, Sep. 2012.
DOI: 10.1021/jp3051033.

[77] M. J. Pilling and S. H. Robertson, "Master Equation Models for Chemical Reactions of Importance in Combustion,"
Annual Review of Physical Chemistry, vol. 54, no. 1, pp. 245–275, 2003.
DOI: 10.1146/annurev.physchem.54.011002.103822.

[78] R. Schulz, Y. von Hansen, J. O. Daldrop, J. Kappler, F. Noé, and R. R. Netz, "Collective hydrogen-bond rearrangement dynamics in liquid water,"
The Journal of Chemical Physics, vol. 149, no. 24, p. 244 504, Dec. 2018.
DOI: 10.1063/1.5054267.

[79] B. L. de Groot, X. Daura, A. E. Mark, and H. Grubmüller, "Essential dynamics of reversible peptide folding: Memory-free conformational dynamics governed by internal hydrogen bonds,"
Journal of Molecular Biology, vol. 309, no. 1, pp. 299–313, May 2001.
DOI: 10.1006/jmbi.2001.4655.

[80] F. Noé and E. Rosta, "Markov Models of Molecular Kinetics,"
The Journal of Chemical Physics, vol. 151, no. 19, p. 190 401, Nov. 2019.

DOI: 10.1063/1.5134029.

[81] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for Markov model construction,"
The Journal of Chemical Physics, vol. 139, no. 1, p. 015 102, Jul. 2013.
DOI: 10.1063/1.4811489.

[82] G. R. Bowman, X. Huang, and V. S. Pande, "Using generalized ensemble simulations and Markov state models to identify conformational states,"
Methods, vol. 49, no. 2, pp. 197–201, Oct. 2009.
DOI: 10.1016/j.ymeth.2009.04.013.

[83] S. Krivov and M. Karplus, "Hidden complexity of free energy surfaces for peptide (protein) folding,"
Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. 41, pp. 14 766–14 770, 2004.
DOI: 10.1073/pnas.0406234101.

[84] G. Berezovska, D. Prada-Gracia, S. Mostarda, and F. Rao, "Accounting for the kinetics in order parameter analysis: Lessons from theoretical models and a disordered peptide,"
The Journal of Chemical Physics, vol. 137, no. 19, p. 194 101, Nov. 2012.
DOI: 10.1063/1.4764868.

[85] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, "A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo,"
Journal of Computational Physics, vol. 151, no. 1, pp. 146–168, May 1999.
DOI: 10.1006/jcph.1999.6231.

[86] S.-H. Chong and S. Ham, "Examining a Thermodynamic Order Parameter of Protein Folding,"
Scientific Reports, vol. 8, no. 1, p. 7148, May 2018.
DOI: 10.1038/s41598-018-25406-8.

[87] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm,"
Expert Systems with Applications, vol. 40, no. 1, pp. 200–210, Jan. 2013.
DOI: 10.1016/j.eswa.2012.07.021.

[88] R. P. D. Bank,
PDB Statistics: PDB Data Distribution by Atom Count,
https://www.rcsb.org/stats/distribution-atom-count.

[89] C. R. Schwantes and V. S. Pande, "Improvements in Markov State Model construction reveal many non-native interactions in the folding of NTL9,"
Journal of Chemical Theory and Computation, vol. 9, no. 4, pp. 2000–2009, 2013.
DOI: 10.1021/ct300878a.

[90] F. Noé and C. Clementi, "Kinetic Distance and Kinetic Maps from Molecular Dynamics Simulation,"
Journal of Chemical Theory and Computation, vol. 11, no. 10, pp. 5002–5011, Oct. 2015.
DOI: 10.1021/acs.jctc.5b00553.

[91] F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé, "Variational Approach to Molecular Kinetics,"
Journal of Chemical Theory and Computation, vol. 10, no. 4, pp. 1739–1752, Apr. 2014.
DOI: 10.1021/ct4009156.

[92] B. E. Husic, R. T. McGibbon, M. M. Sultan, and V. S. Pande, "Optimized parameter selection reveals trends in Markov state models for protein folding,"
The Journal of Chemical Physics, vol. 145, no. 19, p. 194 103, Nov. 2016.
DOI: 10.1063/1.4967809.

[93] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How Fast-Folding Proteins Fold,"
Science, vol. 334, no. 6055, pp. 517–520, Oct. 2011.
DOI: 10.1126/science.1208351.

[94] R. T. McGibbon and V. S. Pande, "Variational cross-validation of slow dynamical modes in molecular kinetics,"
Journal of Chemical Physics, vol. 142, no. 12, 2015.
DOI: 10.1063/1.4916292.

[95] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection,"
Statistics Surveys, vol. 4, pp. 40–79, 2009.
DOI: 10.1214/09-SS054.

[96] H. Wu, F. Nüske, F. Paul, S. Klus, P. Koltai, and F. Noé, "Variational Koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations,"
The Journal of Chemical Physics, vol. 146, no. 15, p. 154 104, Apr. 2017.
DOI: 10.1063/1.4979344.

[97] H. Wu and F. Noé, "Variational Approach for Learning Markov Processes from Time Series Data,"
Journal of Nonlinear Science, vol. 30, no. 1, pp. 23–66, Feb. 2020.
DOI: 10.1007/s00332-019-09567-y.

[98] A. Mardt, L. Pasquali, H. Wu, and F. Noé, "VAMPnets for deep learning of molecular kinetics,"
Nature Communications, vol. 9, no. 1, pp. 1–14, 2018.
DOI: 10.1038/s41467-017-02388-1.

[99] J. R. Porter, M. I. Zimmerman, and G. R. Bowman, "**Enspara** : Modeling molecular ensembles with scalable data structures and parallel computing,"
The Journal of Chemical Physics, vol. 150, no. 4, p. 044 108, Jan. 2019.

DOI: 10.1063/1.5063794.

[100] K. M. Hart, C. M. W. Ho, S. Dutta, M. L. Gross, and G. R. Bowman, "Modelling proteins' hidden conformations to predict antibiotic resistance,"
Nature Communications, vol. 7, no. 1, p. 12 965, Oct. 2016.
DOI: 10.1038/ncomms12965.

[101] D. J. Wales, "Energy landscapes: Calculating pathways and rates,"
International Reviews in Physical Chemistry, vol. 25, no. 1-2, pp. 237–282, Jan. 2006.
DOI: 10.1080/01442350600676921.

[102] D. Wales,
Energy Landscapes: Applications to Clusters, Biomolecules and Glasses.
Cambridge University Press, 2004.
DOI: 10.1017/CBO9780511721724.

[103] D. A. Evans and D. J. Wales, "The free energy landscape and dynamics of met-enkephalin,"
The Journal of Chemical Physics, vol. 119, no. 18, pp. 9947–9955, Nov. 2003.
DOI: 10.1063/1.1616515.

[104] J. A. Joseph, C. S. Whittleston, and D. J. Wales, "Structure, Thermodynamics, and Folding Pathways for a Tryptophan Zipper as a Function of Local Rigidification,"
Journal of Chemical Theory and Computation, vol. 12, no. 12, pp. 6109–6117, Dec. 2016.
DOI: 10.1021/acs.jctc.6b00734.

[105] J. A. Joseph and D. J. Wales, "Intrinsically Disordered Landscapes for Human CD4 Receptor Peptide,"
The Journal of Physical Chemistry B, vol. 122, no. 50, pp. 11 906–11 921, Dec. 2018.
DOI: 10.1021/acs.jpcb.8b08371.

[106] K. Röder and D. J. Wales, "Transforming the Energy Landscape of a Coiled-Coil Peptide via Point Mutations,"
Journal of Chemical Theory and Computation, vol. 13, no. 3, pp. 1468–1477, Mar. 2017.
DOI: 10.1021/acs.jctc.7b00024.

[107] H. Frauenfelder, S. Sligar, and P. Wolynes, "The Energy Landscapes and Motions of Proteins,"
Science, vol. 254, no. 5038, pp. 1598–1603, 1991.

[108] D. Shalloway, "Macrostates of classical stochastic systems,"
The Journal of Chemical Physics, vol. 105, no. 22, pp. 9986–10 007, Dec. 1996.
DOI: 10.1063/1.472830.

[109] Y.-C. Song, S. Ingram, R. E. Arbon, D. O. Topping, D. R. Glowacki, and J. P. Reid, "Transient cavity dynamics and divergence from the Stokes–Einstein equation in organic aerosol,"
Chemical Science, vol. 11, no. 11, pp. 2999–3006, 2020.
DOI: 10.1039/C9SC06228A.

[110] Stephen Ingram,
"Causes, Magnitudes and Atmospheric Implications of Kinetically Limited Diffusion in Organic Aerosol,"
Ph.D. dissertation, University of Bristol, Dec. 2019.

[111] S. Ingram, C. Cai, Y.-C. Song, D. R. Glowacki, D. O. Topping, S. O'Meara, and J. P. Reid, "Characterising the evaporation kinetics of water and semi-volatile organic compounds from viscous multicomponent organic aerosol particles,"
Physical Chemistry Chemical Physics, vol. 19, no. 47, pp. 31 634–31 646, 2017.
DOI: 10.1039/C7CP05172G.

[112] "Stop denying the risks of air pollution,"
Nature, vol. 568, no. 7753, pp. 433–433, Apr. 2019.
DOI: 10.1038/d41586-019-01234-2.

[113] P. Irvine, K. Emanuel, J. He, L. W. Horowitz, G. Vecchi, and D. Keith, "Halving warming with idealized solar geoengineering moderates key climate hazards,"
Nature Climate Change, vol. 9, no. 4, pp. 295–299, Apr. 2019.
DOI: 10.1038/s41558-019-0398-8.

[114] D. K. Farmer, C. D. Cappa, and S. M. Kreidenweis, "Atmospheric Processes and Their Controlling Influence on Cloud Condensation Nuclei Activity,"
Chemical Reviews, vol. 115, no. 10, pp. 4199–4217, May 2015.
DOI: 10.1021/cr5006292.

[115] C. L. McConnell, E. J. Highwood, H. Coe, P. Formenti, B. Anderson, S. Osborne, S. Nava, K. Desboeufs, G. Chen, and M. a. J. Harrison, "Seasonal variations of the physical and optical characteristics of Saharan dust: Results from the Dust Outflow and Deposition to the Ocean (DODO) experiment,"
Journal of Geophysical Research: Atmospheres, vol. 113, no. D14, 2008.
DOI: 10.1029/2007JD009606.

[116] V. Varutbangkul, F. J. Brechtel, R. Bahreini, N. L. Ng, M. D. Keywood, J. H. Kroll, R. C. Flagan, J. H. Seinfeld, A. Lee, and A. H. Goldstein, "Hygroscopicity of secondary organic aerosols formed by oxidation of cycloalkenes, monoterpenes, sesquiterpenes, and related compounds,"
Atmospheric Chemistry and Physics, vol. 6, no. 9, pp. 2367–2388, Jun. 2006.
DOI: 10.5194/acp-6-2367-2006.

[117] S. S. Steimer, U. K. Krieger, Y.-F. Te, D. M. Lienhard, A. J. Huisman, B. P. Luo, M. Ammann, and T. Peter, "Electrodynamic balance measurements of thermodynamic, kinetic, and optical aerosol properties inaccessible to bulk methods,"
Atmospheric Measurement Techniques, vol. 8, no. 6, pp. 2397–2408, Jun. 2015.
DOI: 10.5194/amt-8-2397-2015.

[118] I. N. Tang and H. R. Munkelwitz, "Simultaneous Determination of Refractive Index and Density of an Evaporating Aqueous Solution Droplet,"
Aerosol Science and Technology, vol. 15, no. 3, pp. 201–207, Jan. 1991.
DOI: 10.1080/02786829108959527.

[119] K. A. Dill, S. Bromberg, and D. Stigter,
Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience.
Garland Science, 2010.

[120] R. M. Power, S. H. Simpson, J. P. Reid, and A. J. Hudson, "The transition from liquid to solid-like behaviour in ultrahigh viscosity aerosol particles,"
Chemical Science, vol. 4, no. 6, pp. 2597–2604, May 2013.
DOI: 10.1039/C3SC50682G.

[121] H. C. Price, B. J. Murray, J. Mattsson, D. O'Sullivan, T. W. Wilson, K. J. Baustian, and L. G. Benning, "Quantifying water diffusion in high-viscosity and glassy aqueous solutions using a Raman isotope tracer method,"
Atmospheric Chemistry and Physics, vol. 14, no. 8, pp. 3817–3830, Apr. 2014.
DOI: 10.5194/acp-14-3817-2014.

[122] V. Molinero and W. Goddard, "Microscopic mechanism of water diffusion in glucose glasses,"
Physical Review Letters, vol. 95, no. 4, p. 045 701, Jul. 2005.
DOI: 10.1103/PhysRevLett.95.045701.

[123] H. C. Price, J. Mattsson, Y. Zhang, A. K. Bertram, J. F. Davies, J. W. Grayson, S. T. Martin, D. O'Sullivan, J. P. Reid, A. M. J. Rickards, and B. J. Murray, "Water diffusion in atmospherically relevant alpha-pinene secondary organic material,"
Chemical Science, vol. 6, no. 8, pp. 4876–4883, 2015.
DOI: 10.1039/C5SC00685F.

[124] Y. Chenyakin, D. A. Ullmann, E. Evoy, L. Renbaum-Wolff, S. Kamal, and A. K. Bertram, "Diffusion coefficients of organic molecules in sucrose-water solutions and comparison with Stokes-Einstein predictions,"
Atmospheric Chemistry and Physics, vol. 17, pp. 2423–2435, 2017.
DOI: 10.5194/acp-2016-740.

[125] K. R. Harris, "The fractional Stokes-Einstein equation: Application to Lennard-Jones, molecular, and ionic liquids,"
Journal of Chemical Physics, vol. 131, no. 5, p. 054 503, 2009.
DOI: 10.1063/1.3183951.

[126] H. C. Price, J. Mattsson, and B. J. Murray, "Sucrose diffusion in aqueous solution,"
Physical Chemistry Chemical Physics, vol. 18, no. 28, pp. 19 207–19 216, 2016.
DOI: 10.1039/C6CP03238A.

[127] F. Fernandez-Alonso, F. J. Bermejo, S. E. McLain, J. F. Turner, J. J. Molaison, and K. W. Herwig, "Observation of fractional Stokes-Einstein behavior in the simplest hydrogen-bonded liquid," *Physical Review Letters*, vol. 98, no. 7, p. 077 801, 2007.
DOI: 10.1103/PhysRevLett.98.077801.

[128] E. J. Saltzman and K. S. Schweizer, "Activated hopping and dynamical fluctuation effects in hard sphere suspensions and fluids,"
The Journal of Chemical Physics, vol. 125, no. 4, p. 044 509, Jul. 2006.
DOI: 10.1063/1.2217739.

[129] D. L. Bones, J. P. Reid, D. M. Lienhard, and U. K. Krieger, "Comparing the mechanism of water condensation and evaporation in glassy aerosol.,"
Proc. Nat. Acad. Sci., vol. 109, no. 29, pp. 11 613–8, Jul. 2012.
DOI: 10.1073/pnas.1200691109.

[130] M. K. Scherer, B. E. Husic, M. Hoffmann, F. Paul, H. Wu, and F. Noé, "Variational selection of features for molecular kinetics,"
The Journal of Chemical Physics, vol. 150, no. 19, p. 194 108, May 2019.
DOI: 10.1063/1.5083040.

[131] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient Global Optimization of Expensive Black-Box Functions,"
Journal of Global Optimization, vol. 13, no. 4, pp. 455–492, Dec. 1998.
DOI: 10.1023/A:1008306431147.

[132] M. Feurer and F. Hutter,
"Hyperparameter optimization,"
In *Automated Machine Learning: Methods, Systems, Challenges*,
F. Hutter, L. Kotthoff, and J. Vanschoren, Eds.
Springer International Publishing, 2019,
Pp. 3–33.
DOI: 10.1007/978-3-030-05318-5_1.

[133] F. Hutter, H. H. Hoos, and K. Leyton-Brown,
"Sequential model-based optimization for general algorithm configuration,"
In *Learning and Intelligent Optimization*,
C. A. C. Coello, Ed.,
Springer Berlin Heidelberg, 2011,
Pp. 507–523.
DOI: 10.1007/978-3-642-25566-3_40.

[134] J. Snoek, H. Larochelle, and R. P. Adams,
"Practical bayesian optimization of machine learning algorithms,"
In *Advances in Neural Information Processing Systems 25*,

F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds.,
Curran Associates, Inc., 2012,
Pp. 2951–2959.
[Online]. Available: https://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf.

[135] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl,
"Algorithms for hyper-parameter optimization,"
In *Advances in Neural Information Processing Systems 24*,
J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds.,
Curran Associates, Inc., 2011,
Pp. 2546–2554.
[Online]. Available: http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf.

[136] J. Bergstra, D. Yamins, and D. Cox,
"Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures,"
In *Proceedings of the 30th International Conference on Machine Learning*,
S. Dasgupta and D. McAllester, Eds.,
Vol. 28, PMLR, Jun. 2013,
Pp. 115–123.
[Online]. Available: http://proceedings.mlr.press/v28/bergstra13.pdf.

[137] E. Brochu, V. M. Cora, and N. de Freitas,
A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning,
2010.
arXiv: 1012.2599 [cs.LG].

[138] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the Human Out of the Loop: A Review of Bayesian Optimization,"
Proceedings of the IEEE, vol. 104, no. 1, pp. 148–175, Jan. 2016.
DOI: 10.1109/JPROC.2015.2494218.

[139] J. Bergstra, D. Yamins, and D. D. Cox,
"Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms,"
In *Proceedings of the 12th Python in Science Conference*,
S. van der Walt, J. Millman, and K. Huff, Eds.,
2013,
Pp. 13–19.

DOI: 10.25080/Majora-8b375195-003.

[140] R. Martinez-cantin, "BayesOpt : A Bayesian Optimization Library for Nonlinear Optimization, Experimental Design and Bandits,"
Journal of Machine Learning Research, vol. 15, pp. 3915–3919, 2014.
[Online]. Available: https://jmlr.csail.mit.edu/papers/volume15/loh14a/loh14a.pdf.

[141] M. A. Gelbart, J. Snoek, and R. P. Adams,
"Bayesian optimization with unknown constraints,"
In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*,
N. L. Zhang and J. Tian, Eds.,
AUAI Press, Jul. 2014,
Pp. 250–259.
[Online]. Available: http://auai.org/uai2014/proceedings/individuals/107.pdf.

[142] J. Snoek,
"Abstract Bayesian Optimization and Semiparametric Models with Applications to Assistive Technology,"
Ph.D. dissertation, 2013.
[Online]. Available: https://tspace.library.utoronto.ca/bitstream/1807/43732/1/Snoek_Jasper_R_201311_PhD_thesis.pdf.

[143] J. Snoek, K. Swersky, R. S. Zemel, and R. P. Adams,
Input warping for bayesian optimization of non-stationary functions,
2014.
arXiv: 1402.0929 [stat.ML].

[144] K. Swersky, J. Snoek, and R. P. Adams,
"Multi-task bayesian optimization,"
In *Advances in Neural Information Processing Systems 26*,
C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds.,
Curran Associates, Inc., 2013,
Pp. 2004–2012.
[Online]. Available: http://papers.nips.cc/paper/5086-multi-task-bayesian-optimization.pdf.

[145] The GPyOpt authors,
GPyOpt: A Bayesian optimization framework in Python,
2016.
[Online]. Available: http://github.com/SheffieldML/GPyOpt.

[146] K. Kandasamy, K. R. Vysyaraju, W. Neiswanger, B. Paria, C. R. Collins, J. Schneider, B. Poczos, and E. P. Xing, "Tuning hyperparameters without grad students: Scalable and robust bayesian optimisation with Dragonfly,"
Journal of Machine Learning Research, vol. 21, no. 81, pp. 1–27, 2020.
[Online]. Available: https://jmlr.csail.mit.edu/papers/volume21/18-223/18-223.pdf.

[147] J. Liu, S. Tripathi, U. Kurup, and M. Shah,
"Auptimizer - an Extensible, Open-Source Framework for Hyperparameter Tuning,"
In *2019 IEEE International Conference on Big Data*,
Dec. 2019,
Pp. 339–348.
DOI: 10.1109/BigData47090.2019.9006330.

[148] R. T. McGibbon, C. X. Hernández, M. P. Harrigan, S. Kearnes, M. M. Sultan, S. Jastrzebski, B. E. Husic, and V. S. Pande, "Osprey: Hyperparameter Optimization for Machine Learning,"
Journal of Open Source Software, vol. 1, no. 5, p. 34, Sep. 2016.
DOI: 10.21105/joss.00034.

[149] C. E. Rasmussen and C. Williams K. I.,
Gaussian Processes for Machine Learning,
First edition.
The MIT press, 2006.

[150] J. Friedman, T. Hastie, and R. Tibshirani,
The Elements of Statistical Learning,
Second edition.
Springer-Verlag New York, 2009.
DOI: 10.1007/978-0-387-84858-7.

[151] M. E. Karpen, D. J. Tobias, and C. L. Brooks III, "Statistical clustering techniques for the analysis of long molecular dynamics trajectories: Analysis of 2.2-ns trajectories of YPGDV,"
Biochemistry, vol. 32, no. 2, pp. 412–420, 1993.

[152] G. A. Carpenter and S. Grossberg, "ART 2: Self-organization of stable category recognition codes for analog input patterns,"
Applied Optics, vol. 26, no. 23, pp. 4919–4930, Dec. 1987.
DOI: 10.1364/AO.26.004919.

[153] J. Wei and J. C. W. Kuo, "Lumping Analysis in Monomolecular Reaction Systems. Analysis of the Exactly Lumpable System,"
Industrial & Engineering Chemistry Fundamentals, vol. 8, no. 1, pp. 114–123, Feb. 1969.
DOI: 10.1021/i160029a019.

[154] J. C. W. Kuo and J. Wei, "Lumping analysis in monomolecular reaction systems. analysis of approximately lumpable system,"
Industrial & Engineering Chemistry Fundamentals, vol. 8, no. 1, pp. 124–133, 1969.
DOI: 10.1021/i160029a020.

[155] G. Hummer and A. Szabo, "Optimal Dimensionality Reduction of Multistate Kinetic and Markov-State Models,"
The Journal of Physical Chemistry B, vol. 119, no. 29, pp. 9029–9037, Jul. 2015.
DOI: 10.1021/jp508375q.

[156] A. Kells, V. Koskin, E. Rosta, and A. Annibale, "Correlation functions, mean first passage times, and the Kemeny constant,"
The Journal of Chemical Physics, vol. 152, no. 10, p. 104 108, Mar. 2020.
DOI: 10.1063/1.5143504.

[157] L. Martini, A. Kells, R. Covino, G. Hummer, N.-V. Buchete, and E. Rosta, "Variational Identification of Markovian Transition States,"
Physical Review X, vol. 7, no. 3, p. 031 060, Sep. 2017.
DOI: 10.1103/PhysRevX.7.031060.

[158] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte, "Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains,"
Linear Algebra and its Applications, vol. 315, no. 1, pp. 39–59, Aug. 2000.
DOI: 10.1016/S0024-3795(00)00095-1.

[159] P. Deuflhard and M. Weber, "Robust Perron cluster analysis in conformation dynamics,"
Linear Algebra and its Applications, vol. 398, pp. 161–184, Mar. 2005.
DOI: 10.1016/j.laa.2004.10.026.

[160] N. Plattner and F. Noé, "Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models,"
Nature Communications, vol. 6, no. 1, p. 7653, Jul. 2015.
DOI: 10.1038/ncomms8653.

[161] B. Islam, P. Stadlbauer, A. Gil-Ley, G. Pérez-Hernández, S. Haider, S. Neidle, G. Bussi, P. Banas, M. Otyepka, and J. Sponer, "Exploring the Dynamics of Propeller Loops in Human Telomeric DNA Quadruplexes Using Atomistic Simulations,"
Journal of Chemical Theory and Computation, vol. 13, no. 6, pp. 2458–2480, Jun. 2017.
DOI: 10.1021/acs.jctc.7b00226.

[162] Y. Yao, R. Z. Cui, G. R. Bowman, D.-A. Silva, J. Sun, and X. Huang, "Hierarchical Nyström methods for constructing Markov state models for conformational dynamics,"
The Journal of Chemical Physics, vol. 138, no. 17, p. 174 106, May 2013.
DOI: 10.1063/1.4802007.

[163] G. R. Bowman, "Improved coarse-graining of Markov state models via explicit consideration of statistical uncertainty,"
The Journal of Chemical Physics, vol. 137, no. 13, p. 134 111, Oct. 2012.
DOI: 10.1063/1.4755751.

[164] A. Jain and G. Stock, "Identifying Metastable States of Folding Proteins,"
Journal of Chemical Theory and Computation, vol. 8, no. 10, pp. 3810–3819, Oct. 2012.
DOI: 10.1021/ct300077q.

[165] S. Orioli and P. Faccioli, "Dimensional reduction of Markov state models from renormalization group theory,"
The Journal of Chemical Physics, vol. 145, no. 12, p. 124 120, Sep. 2016.
DOI: 10.1063/1.4963196.

[166] B. E. Husic, K. A. McKiernan, H. K. Wayment-Steele, M. M. Sultan, and V. S. Pande, "A Minimum Variance Clustering Approach Produces Robust and Interpretable Coarse-Grained Models,"
Journal of Chemical Theory and Computation, vol. 14, no. 2, pp. 1071–1082, Feb. 2018.
DOI: 10.1021/acs.jctc.7b01004.

[167] H. Wu, J.-H. Prinz, and F. Noé, "Projected metastable Markov processes and their estimation with observable operator models,"
The Journal of Chemical Physics, vol. 143, no. 14, p. 144 101, Oct. 2015.
DOI: 10.1063/1.4932406.

[168] G. R. Bowman, L. Meng, and X. Huang, "Quantitative comparison of alternative methods for coarse-graining biological networks,"
The Journal of Chemical Physics, vol. 139, no. 12, p. 121 905, Jul. 2013.
DOI: 10.1063/1.4812768.

[169] F. Noé, H. Wu, J.-H. Prinz, and N. Plattner, "Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules,"
The Journal of Chemical Physics, vol. 139, no. 18, p. 184 114, Nov. 2013.
DOI: 10.1063/1.4828816.

[170] L. R. Welch, "Hidden markov models and the Baum-Welch algorithm,"
IEEE Information Theory Society Newsletter, vol. 53, no. 4, pp. 10–13, 2003.
[Online]. Available: https://www.itsoc.org/publications/newsletters/itNL1203.pdf.

[171] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition,"
Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, Feb. 1989.
DOI: 10.1109/5.18626.

[172] J. Pohle, R. Langrock, F. M. van Beest, and N. M. Schmidt, "Selecting the Number of States in Hidden Markov Models: Pragmatic Solutions Illustrated Using Animal Movement," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 22, no. 3, pp. 270–293, 2017.
DOI: 10.1007/s13253-017-0283-8.

[173] J. Mondal, N. Ahalawat, S. Pandit, L. E. Kay, and P. Vallurupalli, "Atomic resolution mechanism of ligand binding to a solvent inaccessible cavity in T4 lysozyme," *PLOS Computational Biology*, vol. 14, no. 5, e1006180, May 2018.
DOI: 10.1371/journal.pcbi.1006180.

[174] N. Plattner, S. Doerr, G. De Fabritiis, and F. Noé, "Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling," *Nature Chemistry*, vol. 9, no. 10, pp. 1005–1011, Oct. 2017.
DOI: 10.1038/nchem.2785.

[175] X. Pan and S. D. Schwartz, "Conformational Heterogeneity in the Michaelis Complex of Lactate Dehydrogenase: An Analysis of Vibrational Spectroscopy Using Markov and Hidden Markov Models," *The Journal of Physical Chemistry B*, vol. 120, no. 27, pp. 6612–6620, Jul. 2016.
DOI: 10.1021/acs.jpcb.6b05119.

[176] J. Juárez-Jiménez, A. A. Gupta, G. Karunanithy, A. S. J. S. Mey, C. Georgiou, H. Ioannidis, A. D. Simone, P. N. Barlow, A. N. Hulme, M. D. Walkinshaw, A. J. Baldwin, and J. Michel, "Dynamic design: Manipulation of millisecond timescale motions on the energy landscape of cyclophilin A," *Chemical Science*, vol. 11, no. 10, pp. 2670–2680, 2020.
DOI: 10.1039/C9SC04696H.

[177] F. Wang, H. Zhou, X. Wang, and P. Tao, "Dynamical Behavior of β-Lactamases and Penicillin-Binding Proteins in Different Functional States and Its Potential Role in Evolution," *Entropy*, vol. 21, no. 11, p. 1130, Nov. 2019.
DOI: 10.3390/e21111130.

[178] "Fast-Folding Pathways of the Thrombin-Binding Aptamer G-Quadruplex Revealed by a Markov State Model," *Biophysical Journal*, vol. 114, no. 7, pp. 1529–1538, Apr. 2018.
DOI: 10.1016/j.bpj.2018.02.021.

[179] J. M. Remington, "Fluorescence quenching in 2-aminopurine-labeled model DNA systems," 2019, Pp. 1–214.

[Online]. Available: https://scholarworks.montana.edu/xmlui/bitstream/handle/1/15563/RemingtonJFluorescenceQuenching2019.pdf.

[180] C. Curado-Carballada, F. Feixas, J. Iglesias-Fernández, and S. Osuna, "Hidden Conformations in Aspergillus niger Monoamine Oxidase are Key for Catalytic Efficiency,"
Angewandte Chemie International Edition, vol. 58, no. 10, pp. 3097–3101, 2019.
DOI: 10.1002/anie.201812532.

[181] S. Furini and C. Domene, "Ion-triggered selectivity in bacterial sodium channels,"
Proceedings of the National Academy of Sciences, vol. 115, no. 21, pp. 5450–5455, May 2018.
DOI: 10.1073/pnas.1722516115.

[182] W. Yang, B. T. Riley, X. Lei, B. T. Porebski, I. Kass, A. M. Buckle, and S. McGowan, "Mapping the Pathway and Dynamics of Bestatin Inhibition of the Plasmodium falciparum M1 Aminopeptidase PfA-M1,"
ChemMedChem, vol. 13, no. 23, pp. 2504–2513, Dec. 2018.
DOI: 10.1002/cmdc.201800563.

[183] N. Ahalawat and J. Mondal, "Mapping the Substrate Recognition Pathway in Cytochrome P450,"
Journal of the American Chemical Society, vol. 140, no. 50, pp. 17 743–17 752, Dec. 2018.
DOI: 10.1021/jacs.8b10840.

[184] O. I. Olaposi, N. Oyekanmi, A. A. Ojo, and G. O. Eniafe, "Membrane-Bound Transcription Factor Site-1 Protease in PF429242 Bound State: Computational Kinetics and Dynamics of Reversible Binding,"
Drug Research, vol. 69, no. 12, pp. 643–649, Nov. 2019.
DOI: 10.1055/a-0807-8640.

[185] J. Xiao and F. R. Salsbury, "Na + -binding modes involved in thrombin's allosteric response as revealed by molecular dynamics simulations, correlation networks and Markov modeling,"
Physical Chemistry Chemical Physics, vol. 21, no. 8, pp. 4320–4330, 2019.
DOI: 10.1039/C8CP07293K.

[186] S. M. Hanson, G. Georghiou, M. K. Thakur, W. T. Miller, J. S. Rest, J. D. Chodera, and M. A. Seeliger, "What Makes a Kinase Promiscuous for Inhibitors?"
Cell Chemical Biology, vol. 26, no. 3, 390–399.e5, Mar. 2019.
DOI: 10.1016/j.chembiol.2018.11.005.

[187] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set,"
Psychometrika, vol. 50, no. 2, pp. 159–179, Jun. 1985.
DOI: 10.1007/BF02294245.

[188] S. Bacallado, J. D. Chodera, and V. Pande, "Bayesian comparison of Markov models of molecular dynamics with detailed balance constraint,"
The Journal of Chemical Physics, vol. 131, no. 4, p. 045 106, Jul. 2009.
DOI: 10.1063/1.3192309.

[189] R. E. Kass and A. E. Raftery, "Bayes Factors,"
Journal of the American Statistical Association, vol. 90, no. 430, pp. 773–795, Jun. 1995.
DOI: 10.1080/01621459.1995.10476572.

[190] G. J. McLachlan and D. Peel,
Finite Mixture Models.
John Wiley & Sons, 2000.

[191] H. Akaike,
"Information Theory and an Extension of the Maximum Likelihood Principle,"
In *Selected Papers of Hirotugu Akaike*,
E. Parzen, K. Tanabe, and G. Kitagawa, Eds.,
Springer, 1998,
Pp. 199–213.
DOI: 10.1007/978-1-4612-1694-0_15.

[192] G. Schwarz, "Estimating the Dimension of a Model,"
The Annals of Statistics, vol. 6, no. 2, pp. 461–464, Mar. 1978.
DOI: 10.1214/aos/1176344136.

[193] G. Celeux and J. B. Durand, "Selecting hidden Markov model state number with cross-validated likelihood,"
Computational Statistics, vol. 23, no. 4, pp. 541–564, 2008.
DOI: 10.1007/s00180-007-0097-1.

[194] R. T. McGibbon, C. R. Schwantes, and V. S. Pande, "Statistical Model Selection for Markov Models of Biomolecular Dynamics,"
The Journal of Physical Chemistry B, vol. 118, no. 24, pp. 6475–6481, Jun. 2014.
DOI: 10.1021/jp411822r.

[195] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood,"
IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 7, pp. 719–725, Jul. 2000.
DOI: 10.1109/34.865189.

[196] A. Brochado and F. V. Martins, "Determining the number of components in mixture regression models: An experimental design,"
Determining the number of components in mixture regression models: an experimental design, no. 2, pp. 1465–1474, 2020.

[197] L. Masgrau, A. Roujeinikova, L. O. Johannissen, P. Hothi, J. Basran, K. E. Ranaghan, A. J. Mulholland, M. J. Sutcliffe, N. S. Scrutton, and D. Leys, "Atomic Description of an Enzyme Reaction Dominated by Proton Tunneling,"
Science, vol. 312, no. 5771, pp. 237–241, Apr. 2006.
DOI: 10.1126/science.1126002.

[198] D. Antoniou and S. D. Schwartz, "Large kinetic isotope effects in enzymatic proton transfer and the role of substrate oscillations,"
Proceedings of the National Academy of Sciences, vol. 94, no. 23, pp. 12 360–12 365, Nov. 1997.
DOI: 10.1073/pnas.94.23.12360.

[199] J. P. Klinman,
"Chapter 6 beyond tunnelling corrections: Full tunnelling models for enzymatic C – H activation reactions,"
In *Quantum Tunnelling in Enzyme-Catalysed Reactions*,
The Royal Society of Chemistry, 2009,
Pp. 132–160.
DOI: 10.1039/9781847559975-00132.

[200] J. Basran, S. Patel, M. J. Sutcliffe, and N. S. Scrutton, "Importance of barrier shape in enzyme-catalyzed reactions: Assisted hydrogen tunneling in tryptophan tryptophylquinone-dependent amine dehydrogenase,"
Journal of Biological Chemistry, vol. 276, no. 9, pp. 6234–6242, Feb. 2001.
DOI: 10.1074/jbc.M008141200.

[201] H. B. Brooks, L. H. Jones, and V. L. Davidson, "Deuterium kinetic isotope effect and stopped-flow kinetic studies of the quinoprotein methylamine dehydrogenase,"
Biochemistry, vol. 32, no. 10, pp. 2725–2729, Mar. 1993.
DOI: 10.1021/bi00061a034.

[202] J. Basran, M. J. Sutcliffe, and N. S. Scrutton, "Enzymatic H-Transfer Requires Vibration-Driven Extreme Tunneling,"
Biochemistry, vol. 38, no. 10, pp. 3218–3222, Mar. 1999.
DOI: 10.1021/bi982719d.

[203] M. H. Glickman, J. S. Wiseman, and J. P. Klinman, "Extremely Large Isotope Effects in the Soybean Lipoxygenase-Linoleic Acid Reaction,"
Journal of the American Chemical Society, vol. 116, no. 2, pp. 793–794, Jan. 1994.
DOI: 10.1021/ja00081a060.

[204] M. J. Knapp, K. Rickert, and J. P. Klinman, "Temperature-Dependent Isotope Effects in Soybean Lipoxygenase-1: Correlating Hydrogen Tunneling with Protein Dynamics,"
Journal of the American Chemical Society, vol. 124, no. 15, pp. 3865–3874, Apr. 2002.
DOI: 10.1021/ja012205t.

[205] R. S. Sikorski, L. Wang, K. A. Markham, P. T. R. Rajagopalan, S. J. Benkovic, and A. Kohen, "Tunneling and Coupled Motion in the Escherichia coli Dihydrofolate Reductase Catalysis," *Journal of the American Chemical Society*, vol. 126, no. 15, pp. 4778–4779, Apr. 2004. DOI: 10.1021/ja031683w.

[206] E. J. Loveridge, L.-H. Tey, and R. K. Allemann, "Solvent Effects on Catalysis by Escherichia coli Dihydrofolate Reductase," *Journal of the American Chemical Society*, vol. 132, no. 3, pp. 1137–1143, Jan. 2010. DOI: 10.1021/ja909353c.

[207] J. Pu, J. Gao, and D. G. Truhlar, "Multidimensional Tunneling, Recrossing, and the Transmission Coefficient for Enzymatic Reactions," *Chemical Reviews*, vol. 106, no. 8, pp. 3140–3169, Aug. 2006. DOI: 10.1021/cr050308e.

[208] J. P. Klinman and A. Kohen, "Hydrogen Tunneling Links Protein Dynamics to Enzyme Catalysis," *Annual Review of Biochemistry*, vol. 82, no. 1, pp. 471–496, 2013. DOI: 10.1146/annurev-biochem-051710-133623.

[209] M. Garcia-Viloca, J. Gao, M. Karplus, and D. G. Truhlar, "How Enzymes Work: Analysis by Modern Rate Theory and Computer Simulations," *Science*, vol. 303, no. 5655, pp. 186–195, Jan. 2004. DOI: 10.1126/science.1088172.

[210] L. Masgrau, J. Basran, P. Hothi, M. J. Sutcliffe, and N. S. Scrutton, "Hydrogen tunneling in quinoproteins," *Archives of biochemistry and biophysics*, vol. 428, no. 1, pp. 41–51, 2004. DOI: 10.1016/j.abb.2004.03.013.

[211] W. J. Bruno and W. Bialek, "Vibrationally enhanced tunneling as a mechanism for enzymatic hydrogen transfer.," *Biophysical journal*, vol. 63, no. 3, pp. 689–699, 1992. DOI: 10.1016/S0006-3495(92)81654-5.

[212] D. Borgis and J. T. Hynes, "Curve Crossing Formulation for Proton Transfer Reactions in Solution," *The Journal of Physical Chemistry*, vol. 100, no. 4, pp. 1118–1128, Jan. 1996. DOI: 10.1021/jp9522324.

[213] R. A. Marcus and N. Sutin, "Electron transfers in chemistry and biology," *Biochimica et Biophysica Acta (BBA) - Reviews on Bioenergetics*, vol. 811, no. 3, pp. 265–322, Aug. 1985. DOI: 10.1016/0304-4173(85)90014-X.

[214] A. Kuznetsov and J. Ulstrup, "Proton and hydrogen atom tunnelling in hydrolytic and redox enzyme catalysis,"
Canadian Journal of Chemistry, vol. 77, no. 5-6, pp. 1085–1096, 1999.
DOI: 10.1139/cjc-77-5-6-1085.

[215] R. K. Allemann and N. S. Scrutton, Eds.,
Quantum Tunnelling in Enzyme-Catalysed Reactions.
The Royal Society of Chemistry, 2009,
P001–P385.
DOI: 10.1039/9781847559975. [Online]. Available: http://dx.doi.org/10.1039/9781847559975.

[216] L. O. Johannissen, S. Hay, N. S. Scrutton, and M. J. Sutcliffe, "Proton Tunneling in Aromatic Amine Dehydrogenase is Driven by a Short-Range Sub-Picosecond Promoting Vibration: Consistency of Simulation and Theory with Experiment,"
The Journal of Physical Chemistry B, vol. 111, no. 10, pp. 2631–2638, Mar. 2007.
DOI: 10.1021/jp066276w.

[217] L. O. Johannissen, N. S. Scrutton, and M. J. Sutcliffe, "The enzyme aromatic amine dehydrogenase induces a substrate conformation crucial for promoting vibration that significantly reduces the effective potential energy barrier to proton transfer,"
Journal of The Royal Society Interface, vol. 5, no. suppl_3, pp. 225–232, Dec. 2008.
DOI: 10.1098/rsif.2008.0068.focus.

[218] M. J. Sutcliffe, L. Masgrau, A. Roujeinikova, L. O. Johannissen, P. Hothi, J. Basran, K. E. Ranaghan, A. J. Mulholland, D. Leys, and N. S. Scrutton, "Hydrogen tunnelling in enzyme-catalysed H-transfer reactions: Flavoprotein and quinoprotein systems,"
Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 361, no. 1472, pp. 1375–1386, Aug. 2006.
DOI: 10.1098/rstb.2006.1878.

[219] V. Stojković, L. L. Perissinotti, D. Willmer, S. J. Benkovic, and A. Kohen, "Effects of the donor–acceptor distance and dynamics on hydride tunneling in the dihydrofolate reductase catalyzed reaction,"
Journal of the American Chemical Society, vol. 134, no. 3, pp. 1738–1745, 2012.
DOI: 10.1021/ja209425w.

[220] D. R. Glowacki, J. N. Harvey, and A. J. Mulholland, "Protein dynamics and enzyme catalysis: The ghost in the machine?"
Biochemical Society Transactions, vol. 40, no. 3, pp. 515–521, Jun. 2012.
DOI: 10.1042/BST20120047.

[221] D. R. Glowacki, J. N. Harvey, and A. J. Mulholland, "Taking Ockham's razor to enzyme dynamics and catalysis,"

Nature Chemistry, vol. 4, no. 3, pp. 169–176, Mar. 2012.
DOI: 10.1038/nchem.1244.

[222] D. Antoniou and S. D. Schwartz, "Protein Dynamics and Enzymatic Chemical Barrier Passage," *The Journal of Physical Chemistry B*, vol. 115, no. 51, pp. 15 147–15 158, Dec. 2011.
DOI: 10.1021/jp207876k.

[223] A. J. Adamczyk, J. Cao, S. C. L. Kamerlin, and A. Warshel, "Catalysis by dihydrofolate reductase and other enzymes arises from electrostatic preorganization, not conformational motions," *Proceedings of the National Academy of Sciences*, vol. 108, no. 34, pp. 14 115–14 120, Aug. 2011.

[224] A. V. Pisliakov, J. Cao, S. C. L. Kamerlin, and A. Warshel, "Enzyme millisecond conformational dynamics do not catalyze the chemical step.,"
Proceedings of the National Academy of Sciences of the United States of America, vol. 106, no. 41, pp. 17 359–17 364, 2009.
DOI: 10.1073/pnas.0909150106.

[225] J. D. McGeagh, K. E. Ranaghan, and A. J. Mulholland, "Protein dynamics and enzyme catalysis: Insights from simulations,"
Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, vol. 1814, no. 8, pp. 1077–1092, Aug. 2011.
DOI: 10.1016/j.bbapap.2010.12.002.

[226] E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, and D. Kern, "Intrinsic dynamics of an enzyme underlies catalysis,"
Nature, vol. 438, no. 7064, pp. 117–121, Nov. 2005.
DOI: 10.1038/nature04105.

[227] K. A. Henzler-Wildman, V. Thai, M. Lei, M. Ott, M. Wolf-Watz, T. Fenn, E. Pozharski, M. A. Wilson, G. A. Petsko, M. Karplus, C. G. Hübner, and D. Kern, "Intrinsic motions along an enzymatic reaction trajectory,"
Nature, vol. 450, no. 7171, pp. 838–844, Dec. 2007.
DOI: 10.1038/nature06410.

[228] M. H. M. Olsson and A. Warshel, "Solute Solvent Dynamics and Energetics in Enzyme Catalysis: The SN2 Reaction of Dehalogenase as a General Benchmark,"
Journal of the American Chemical Society, vol. 126, no. 46, pp. 15 167–15 179, Nov. 2004.
DOI: 10.1021/ja047151c.

[229] S. C. L. Kamerlin and A. Warshel, "Multiscale modeling of biological functions,"
Physical Chemistry Chemical Physics, vol. 13, no. 22, pp. 10 401–10 411, May 2011.
DOI: 10.1039/C0CP02823A.

[230] W. Wang, S. Cao, L. Zhu, and X. Huang, "Constructing Markov State Models to elucidate the functional conformational changes of complex biomolecules,"

WIREs Computational Molecular Science, vol. 8, no. 1, e1343, 2018.
DOI: 10.1002/wcms.1343.

[231] D. L. Ermak, "A computer simulation of charged particles in solution. I. Technique and equilibrium properties,"
The Journal of Chemical Physics, vol. 62, no. 10, p. 4189, Jan. 1974.
DOI: 10.1063/1.430300.

[232] D. L. Ermak and Y. Yeh, "Equilibrium electrostatic effects on the behavior of polyions in solution: Polyion-mobile ion interaction,"
Chemical Physics Letters, vol. 24, no. 2, pp. 243–248, Jan. 1974.
DOI: 10.1016/0009-2614(74)85442-4.

[233] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling,"
The Journal of Chemical Physics, vol. 126, no. 1, p. 014 101, Jan. 2007.
DOI: 10.1063/1.2408420.

[234] F. Noé, I. Horenko, C. Schütte, and J. C. Smith, "Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states,"
The Journal of Chemical Physics, vol. 126, no. 15, p. 155 102, Apr. 2007.
DOI: 10.1063/1.2714539.

[235] J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill, "Long-Time Protein Folding Dynamics from Short-Time Molecular Dynamics Simulations,"
Multiscale Modeling & Simulation, vol. 5, no. 4, pp. 1214–1226, Jan. 2006.
DOI: 10.1137/06065146X.

[236] S. Kube and M. Weber, "A coarse graining method for the identification of transition rates between molecular conformations,"
The Journal of Chemical Physics, vol. 126, no. 2, p. 024 103, Jan. 2007.
DOI: 10.1063/1.2404953.

[237] N.-V. Buchete and G. Hummer, "Coarse Master Equations for Peptide Folding Dynamics,"
The Journal of Physical Chemistry B, vol. 112, no. 19, pp. 6057–6069, May 2008.
DOI: 10.1021/jp0761665.

[238] N. Stanley, S. Esteban-Martín, and G. De Fabritiis, "Kinetic modulation of a disordered protein domain by phosphorylation,"
Nature Communications, vol. 5, no. 1, p. 5272, Oct. 2014.
DOI: 10.1038/ncomms6272.

[239] T. Zhou and A. Caflisch, "Distribution of Reciprocal of Interatomic Distances: A Fast Structural Metric,"
Journal of Chemical Theory and Computation, vol. 8, no. 8, pp. 2930–2937, Aug. 2012.
DOI: 10.1021/ct3003145.

[240] D. Shukla, A. Peck, and V. S. Pande, "Conformational heterogeneity of the calmodulin binding interface,"
Nature Communications, vol. 7, no. 1, p. 10 910, Apr. 2016.
DOI: 10.1038/ncomms10910.

[241] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, "Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations,"
Proceedings of the National Academy of Sciences, vol. 106, no. 45, pp. 19 011–19 016, 2009.
DOI: 10.1073/pnas.0905466106.

[242] G. R. Bowman, V. A. Voelz, and V. S. Pande, "Atomistic Folding Simulations of the Five-Helix Bundle Protein Λ6-85,"
Journal of the American Chemical Society, vol. 133, no. 4, pp. 664–667, Feb. 2011.
DOI: 10.1021/ja106936n.

[243] S. Muff and A. Caflisch, "Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein,"
Proteins: Structure, Function, and Bioinformatics, vol. 70, no. 4, pp. 1185–1195, 2008.
DOI: 10.1002/prot.21565.

[244] S. Lloyd, "Least squares quantization in PCM,"
IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129–137, Mar. 1982.
DOI: 10.1109/TIT.1982.1056489.

[245] F. Noé and F. Nüske, "A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems,"
Multiscale Modeling & Simulation, vol. 11, no. 2, pp. 635–655, Jan. 2013.
DOI: 10.1137/110858616.

[246] B. E. Husic and V. S. Pande, "Ward Clustering Improves Cross-Validated Markov State Models of Protein Folding,"
Journal of Chemical Theory and Computation, vol. 13, no. 3, pp. 963–967, Mar. 2017.
DOI: 10.1021/acs.jctc.6b01238.

[247] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, "Progress and challenges in the automated construction of Markov state models for full protein systems,"
The Journal of Chemical Physics, vol. 131, no. 12, p. 124 101, Sep. 2009.
DOI: 10.1063/1.3216567.

[248] B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé, "Estimation and uncertainty of reversible Markov models,"
The Journal of Chemical Physics, vol. 143, no. 17, p. 174 101, Nov. 2015.
DOI: 10.1063/1.4934536.

[249] L. Wasserman,
All of Statistics: A Concise Course in Statistical Inference.
Springer-Verlag New York, 2004.
DOI: 10.1007/978-0-387-21736-9.

[250] F. Noé,
"Statistical inefficiency of Markov model count matrices."
[Online]. Available: http://publications.mi.fu-berlin.de/1699/.

[251] R. J. Wilson, "Introduction to graph theory,"
1996.

[252] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin,
Bayesian Data Analysis,
Third edition.
CRC Press, 2014.

[253] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, "PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models,"
Journal of Chemical Theory and Computation, vol. 11, pp. 5525–5542, Oct. 2015.
DOI: 10.1021/acs.jctc.5b00743.

[254] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner, "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC,"
Bayesian Analysis, Jul. 2020.
DOI: 10.1214/20-BA1221.

[255] W. C. Swope, J. W. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B. G. Fitch, R. S. Germain, A. Rayshubski, T. J. C. Ward, Y. Zhestkov, and R. Zhou, "Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 2. Example Applications to Alanine Dipeptide and a β-Hairpin Peptide,"
The Journal of Physical Chemistry B, vol. 108, no. 21, pp. 6582–6594, May 2004.
DOI: 10.1021/jp037422q.

[256] B. Efron and R. J. Tibshirani,
An Introduction to the Bootstrap.
Springer US, 1993.
DOI: 10.1007/978-1-4899-4541-9.

[257] S. Chen, R. P. Wiewiora, F. Meng, N. Babault, A. Ma, W. Yu, K. Qian, H. Hu, H. Zou, J. Wang, S. Fan, G. Blum, F. Pittella-Silva, K. A. Beauchamp, W. Tempel, H. Jiang, K. Chen, R. J. Skene, Y. G. Zheng, P. J. Brown, J. Jin, C. Luo, J. D. Chodera, and M. Luo, "The dynamic conformational landscape of the protein methyltransferase SETD8,"
eLife, vol. 8, S. J. Fleishman, J. Kuriyan, G. R. Bowman, and E. Lindahl, Eds., e45403, May 2019.

DOI: 10.7554/eLife.45403.

[258] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains,"
Annals of Mathematical Statistics, vol. 41, no. 1, pp. 164–171, Feb. 1970.
DOI: 10.1214/aoms/1177697196.

[259] J. D. Chodera, P. Elms, F. Noé, B. Keller, C. M. Kaiser, A. Ewall-Wice, S. Marqusee, C. Bustamante, and N. S. Hinrichs,
Bayesian hidden markov model analysis of single-molecule force spectroscopy: Characterizing kinetics under measurement uncertainty,
2011.
arXiv: 1108.1430 [cond-mat.stat-mech].

[260] M. L. F. Nascimento and E. Dutra Zanotto, "Does viscosity describe the kinetic barrier for crystal growth from the liquidus to the glass transition?"
Journal of Chemical Physics, vol. 133, no. 17, 2010.
DOI: 10.1063/1.3490793.

[261] E. Mikhailov, S. Vlasenko, S. T. Martin, T. Koop, and U. Pöschl, "Amorphous and crystalline aerosol particles interacting with water vapor: Conceptual framework and experimental evidence for restructuring, phase transitions and kinetic limitations,"
Atmospheric Chemistry and Physics, vol. 9, no. 24, pp. 9491–9522, Dec. 2009.
DOI: 10.5194/acp-9-9491-2009.

[262] T. Koop, J. Bookhold, M. Shiraiwa, and U. Pöschl, "Glass transition and phase state of organic compounds: Dependency on molecular properties and implications for secondary organic aerosols in the atmosphere.,"
Physical Chemistry Chemical Physics, vol. 13, no. 43, pp. 19 238–55, Nov. 2011.
DOI: 10.1039/c1cp22617g.

[263] G. Hidy,
Aerosols, an Industrial and Environmental Science.
Academic Press, 1984.

[264] B. C. Hancock and G. Zografi, "Characteristics and significance of the amorphous state in pharmaceutical systems,"
Journal of Pharmaceutical Sciences, vol. 86, no. 1, pp. 1–12, 1997.
DOI: 10.1021/js9601896.

[265] C. Lorenz, H. Hagendorfer, N. Von Goetz, R. Kaegi, R. Gehrig, A. Ulrich, M. Scheringer, and K. Hungerbühler, "Nanosized aerosols from consumer sprays: Experimental analysis and exposure modeling for four commercial products,"
Journal of Nanoparticle Research, vol. 13, no. 8, pp. 3377–3391, 2011.
DOI: 10.1007/s11051-011-0256-8.

[266] A. E. Haddrell, J. F. Davies, R. E. Miles, J. P. Reid, L. A. Dailey, and D. Murnane, "Dynamics of aerosol size during inhalation: Hygroscopic growth of commercial nebulizer formulations," *International Journal of Pharmaceutics*, vol. 463, no. 1, pp. 50–61, 2014.
DOI: 10.1016/j.ijpharm.2013.12.048.

[267] M. Shiraiwa and J. H. Seinfeld, "Equilibration timescale of atmospheric secondary organic aerosol partitioning,"
Geophysical Research Letters, vol. 39, no. 24, Dec. 2012.
DOI: 10.1029/2012GL054008.

[268] H. Mai, M. Shiraiwa, R. C. Flagan, and J. H. Seinfeld, "Under what conditions can equilibrium gas–particle partitioning be expected to hold in the atmosphere?"
Environmental science & technology, vol. 49, no. 19, pp. 11 485–11 491, 2015.

[269] A. M. Maclean, C. L. Butenhoff, J. W. Grayson, K. Barsanti, J. L. Jimenez, and A. K. Bertram, "Mixing times of organic molecules within secondary organic aerosol particles: A global planetary boundary layer perspective,"
Atmospheric Chemistry and Physics, vol. 17, no. 21, pp. 13 037–13 048, Nov. 2017.
DOI: 10.5194/acp-17-13037-2017.

[270] B. Chen, E. E. Sigmund, and W. P. Halperin, "Stokes-Einstein Relation in Supercooled Aqueous Solutions of Glycerol,"
Physical Review Letters, vol. 96, no. 14, p. 145 502, Apr. 2006.
DOI: 10.1103/PhysRevLett.96.145502.

[271] M. Shrestha, Y. Zhang, M. A. Upshur, P. Liu, S. L. Blair, H.-f. Wang, S. A. Nizkorodov, R. J. Thomson, S. T. Martin, and F. M. Geiger, "On surface order and disorder of alpha-Pinene-Derived secondary organic material,"
The Journal of Physical Chemistry A, vol. 119, no. 19, pp. 4609–4617, May 2015.
DOI: 10.1021/jp510780e.

[272] P. Kumar, "Breakdown of the Stokes-Einstein relation in supercooled water,"
Proceedings of the National Academy of Sciences, vol. 103, no. 35, pp. 12 955–12 956, 2006.
DOI: 10.1073/pnas.0605880103.

[273] J. A. T. González, M. P. Longinotti, and H. R. Corti, "Diffusion-viscosity decoupling in supercooled glycerol aqueous solutions,"
Journal of Physical Chemistry B, vol. 119, no. 1, pp. 257–262, 2015.
DOI: 10.1021/jp509055v.

[274] S. Bastelberger, U. K. Krieger, B. Luo, and T. Peter, "Diffusivity measurements of volatile organics in levitated viscous aerosol particles,"
Atmospheric Chemistry and Physics, vol. 17, no. 13, pp. 8453–8471, Jul. 2017.
DOI: 10.5194/acp-17-8453-2017.

[275] N. E. Rothfuss and M. D. Petters, "Influence of functional groups on the viscosity of organic aerosol,"
Environmental science & technology, vol. 51, no. 1, pp. 271–279, 2017.
DOI: 10.1021/acs.est.6b04478.

[276] Y. Zhao, B. G. Trewyn, I. I. Slowing, and V. S. Lin, "Mesoporous silica nanoparticle-based double drug delivery system for glucose-responsive controlled release of insulin and cyclic AMP,"
Journal of the American Chemical Society, vol. 131, no. 24, pp. 8398–8400, 2009.
DOI: 10.1021/ja901831u.

[277] V. R. Sinha and R. Kumria, "Polysaccharides in colon-specific drug delivery,"
International Journal of Pharmaceutics, vol. 224, no. 1-2, pp. 19–38, 2001.
DOI: 10.1016/S0378-5173(01)00720-7.

[278] S. Sastry, J. Nyshadham, and J. Fix, "Recent technological advances in oral drug delivery - a review.,"
Pharmaceutical science & technology today, vol. 3, no. 4, pp. 138–145, 2000.
DOI: 10.1016/S1461-5347(00)00247-9.

[279] J. D. Andya, Y. F. Maa, H. R. Costantino, P. A. Nguyen, N. Dasovich, T. D. Sweeney, C. C. Hsu, and S. J. Shire, "The effect of formulation excipients on protein stability and aerosol performance of spray-dried powders of a recombinant humanized anti-IgE monoclonal antibody,"
Pharmaceutical Research, vol. 16, no. 3, pp. 350–358, 1999.
DOI: 10.1023/A:1018805232453.

[280] R. Vehring, "Pharmaceutical particle engineering via spray drying,"
Pharmaceutical Research, vol. 25, no. 5, pp. 999–1022, 2008.
DOI: 10.1007/s11095-007-9475-1.

[281] K. Mosén, K. Bäckström, K. Thalberg, T. Schaefer, H. G. Kristensen, and A. Axelsson, "Particle formation and capture during spray drying of inhalable particles,"
Pharmaceutical Development and Technology, vol. 9, no. 4, pp. 409–417, 2004.
DOI: 10.1081/PDT-200035795.

[282] Y. Liu, B. Bhandari, and W. Zhou, "Glass transition and enthalpy relaxation of amorphous food saccharides: A review,"
Journal of agricultural and food chemistry, vol. 54, no. 16, pp. 5701–5717, 2006.
DOI: 10.1021/jf060188r.

[283] B. Zobrist, C. Marcolli, D. a. Pedernera, and T. Koop, "Do atmospheric aerosols form glasses?"
Atmospheric Chemistry and Physics, vol. 8, no. 17, pp. 5221–5244, 2008.
DOI: 10.5194/acp-8-5221-2008.

[284] J. P. Reid, A. K. Bertram, D. O. Topping, A. Laskin, S. T. Martin, M. D. Petters, F. D. Pope, and G. Rovelli, "The viscosity of atmospherically relevant organic particles," *Nature Communications*, vol. 9, no. 1, p. 956, Dec. 2018. DOI: 10.1038/s41467-018-03027-z.

[285] Y. C. Song, A. E. Haddrell, B. R. Bzdek, J. P. Reid, T. Bannan, D. O. Topping, C. Percival, and C. Cai, "Measurements and predictions of binary component aerosol particle viscosity," *The Journal of Physical Chemistry A*, vol. 120, no. 41, pp. 8123–8137, Oct. 2016. DOI: 10.1021/acs.jpca.6b07835.

[286] S. Decesari, S. Fuzzi, M. C. Facchini, M. Mircea, L. Emblico, F. Cavalli, W. Maenhaut, X. Chi, G. Schkolnik, A. Falkovich, Y. Rudich, M. Claeys, V. Pashynska, G. Vas, I. Kourtchev, R. Vermeylen, A. Hoffer, M. O. Andreae, E. Tagliavini, F. Moretti, and P. Artaxo, "Characterization of the organic composition of aerosols from Rondônia, Brazil, during the LBA-SMOCC 2002 experiment and its representation through model compounds," *Atmospheric Chemistry and Physics*, vol. 6, no. 2, pp. 375–402, 2006. DOI: 10.5194/acp-6-375-2006.

[287] A. M. J. Rickards, Y.-C. Song, R. E. H. Miles, T. C. Preston, and J. P. Reid, "Variabilities and uncertainties in characterising water transport kinetics in glassy and ultraviscous aerosol," *Physical Chemistry Chemical Physics*, vol. 17, no. 15, pp. 10 059–10 073, 2015. DOI: 10.1039/C4CP05383D.

[288] H. J. Tong, J. P. Reid, D. L. Bones, B. P. Luo, and U. K. Krieger, "Measurements of the timescales for the mass transfer of water in glassy aerosol at low relative humidity and ambient temperature," *Atmospheric Chemistry and Physics*, vol. 11, no. 10, pp. 4739–4754, 2011. DOI: 10.5194/acp-11-4739-2011.

[289] D. M. Lienhard, A. J. Huisman, U. K. Krieger, Y. Rudich, C. Marcolli, B. Luo, D. L. Bones, J. P. Reid, A. T. Lambe, M. R. Canagaratna, *et al.*, "Viscous organic aerosol particles in the upper troposphere: Diffusivity-controlled water uptake and ice nucleation?" *Atmospheric Chemistry and Physics*, vol. 15, no. 23, pp. 13 599–13 613, 2015. DOI: 10.5194/acp-15-13599-2015.

[290] Q. Liu, S. Huang, and Z. Suo, "Brownian motion of molecular probes in supercooled liquids," *Physical Review Letters*, vol. 114, no. 22, p. 224 301, 2015. DOI: 10.1103/PhysRevLett.114.224301.

[291] S. Chen, W. H. Brune, A. T. Lambe, P. Davidovits, and T. B. Onasch, "Modeling organic aerosol from the oxidation of α-pinene in a Potential Aerosol Mass (PAM) chamber," *Atmospheric Chemistry and Physics*, vol. 13, no. 9, pp. 5017–5031, 2013. DOI: 10.5194/acp-13-5017-2013.

[292] V. K. Shante and S. Kirkpatrick, "An introduction to percolation theory,"

Advances in Physics, vol. 20, no. 85, pp. 325–357, May 1971.
DOI: 10.1080/00018737100101261.

[293] M. Shiraiwa, M. Ammann, T. Koop, and U. Poschl, "Gas uptake and chemical aging of semisolid organic aerosol particles,"
Proceedings of the National Academy of Sciences, vol. 108, no. 27, pp. 11 003–11 008, Jul. 2011.
DOI: 10.1073/pnas.1103045108.

[294] R. P. White and J. E. G. Lipson, "Polymer free volume and its connection to the glass transition,"
Macromolecules, vol. 49, no. 11, pp. 3987–4007, Jun. 2016.
DOI: 10.1021/acs.macromol.6b00215.

[295] X. Zhang, R. C. McVay, D. D. Huang, N. F. Dalleska, B. Aumont, R. C. Flagan, and J. H. Seinfeld, "Formation and evolution of molecular products in α-pinene secondary organic aerosol,"
Proceedings of the National Academy of Sciences, vol. 112, no. 46, pp. 14 168–14 173, 2015.
DOI: 10.1073/pnas.1517742112.

[296] R. Krishna and J. Wesselingh, "The Maxwell-Stefan approach to mass transfer,"
Chemical Engineering Science, vol. 52, no. 6, pp. 861–911, Mar. 1997.
DOI: 10.1016/S0009-2509(96)00458-7.

[297] L. Jia and Y. Xu, "Different roles of water in secondary organic aerosol formation from toluene and isoprene,"
Atmospheric Chemistry and Physics, vol. 18, no. 11, pp. 8137–8154, Jun. 2018.
DOI: 10.5194/acp-18-8137-2018.

[298] W. Huang, H. Saathoff, A. Pajunoja, X. Shen, K.-H. Naumann, R. Wagner, A. Virtanen, T. Leisner, and C. Mohr, "α-pinene secondary organic aerosol at low temperature: Chemical composition and implications for particle viscosity,"
Atmospheric Chemistry and Physics, vol. 18, no. 4, pp. 2883–2898, 2018.
DOI: 10.5194/acp-18-2883-2018.

[299] I. Kourtchev, C. Giorio, A. Manninen, E. Wilson, B. Mahon, J. Aalto, M. Kajos, D. Venables, T. Ruuskanen, J. Levula, M. Loponen, S. Connors, N. Harris, D. Zhao, A. Kiendler-Scharr, T. Mentel, Y. Rudich, M. Hallquist, J.-F. Doussin, W. Maenhaut, J. Bäck, T. Petäjä, J. Wenger, M. Kulmala, and M. Kalberer, "Enhanced Volatile Organic Compounds emissions and organic aerosol mass increase the oligomer content of atmospheric aerosols,"
Scientific Reports, vol. 6, no. 1, p. 35 038, Dec. 2016.
DOI: 10.1038/srep35038.

[300] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient Global Optimization of Expensive Black-Box Functions,"
Journal of Global Optimization, vol. 13, no. 4, pp. 455–492, Dec. 1998.
DOI: 10.1023/A:1008306431147.

[301] D. C Montgomery,
Montgomery Design and Analysis of Experiments.
John Wiley, 1997.

[302] D. Simon,
Evolutionary Optimization Algorithms.
John Wiley & Sons, 2013.

[303] J. Kennedy and R. Eberhart,
"Particle swarm optimization,"
In *Proceedings of ICNN'95 - International Conference on Neural Networks*,
Vol. 4, Nov. 1995,
Pp. 1942–1948.
DOI: 10.1109/ICNN.1995.488968.

[304] R. C. Eberhart and Y. Shi,
"Comparison between genetic algorithms and particle swarm optimization,"
In *Evolutionary Programming VII*,
V. W. Porto, N. Saravanan, D. Waagen, and A. E. Eiben, Eds.,
Springer Berlin Heidelberg, 1998,
Pp. 611–616.
DOI: /10.1007/BFb0040812.

[305] N. Hansen,
The CMA evolution strategy: A tutorial,
2016.
arXiv: 1604.00772 [cs.LG].

[306] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization,"
Journal of Machine Learning Research, vol. 13, pp. 281–305, 2012.
[Online]. Available: https://jmlr.csail.mit.edu/papers/volume13/bergstra12a/bergstra12a.pdf.

[307] H. J. Kushner, "A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise,"
Joint Automatic Control Conference, vol. 1, pp. 69–79, 1963.
DOI: 10.1109/JACC.1963.4168566.

[308] J. Močkus,
"On bayesian methods for seeking the extremum,"
In *Optimization Techniques IFIP Technical Conference: Novosibirsk, July 1–7, 1974*,
G. I. Marchuk, Ed.
Springer Berlin Heidelberg, 1975,
Pp. 400–404.

DOI: 10.1007/978-3-662-38527-2_55.

[309] D. R. Jones, "A taxonomy of global optimization methods based on response surfaces," *Journal of global optimization*, vol. 21, no. 4, pp. 345–383, 2001. DOI: 10.1023/A:1012771025575.

[310] N. Srinivas, A. Krause, S. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, J. Fürnkranz and T. Joachims, Eds., Omnipress, Jun. 2010, Pp. 1015–1022. [Online]. Available: https://icml.cc/Conferences/2010/papers/422.pdf.

[311] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media, 2013.

[312] P. Hennig and C. J. Schuler, "Entropy search for information-efficient global optimization," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1809–1837, 2012. [Online]. Available: https://jmlr.csail.mit.edu/papers/volume13/hennig12a/hennig12a.pdf.

[313] J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani, "Predictive entropy search for efficient global optimization of black-box functions," In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2014, Pp. 918–926. [Online]. Available: http://papers.nips.cc/paper/5324-predictive-entropy-search-for-efficient-global-optimization-of-black-box-functions.pdf.

[314] B. Shahriari, Z. Wang, M. W. Hoffman, A. Bouchard-Côté, and N. de Freitas, *An entropy search portfolio for bayesian optimization*, 2015. arXiv: 1406.4625 [stat.ML].

[315] A. J. Dobson, *An introduction to generalized linear models*, Second edition. CRC press, 2002.

[316] J. Quiñonero-Candela and C. E. Rasmussen, "A Unifying View of Sparse Approximate Gaussian Process Regression," *Journal of Machine Learning Research*, vol. 6, no. Dec, pp. 1939–1959, 2005.

[Online]. Available: https://jmlr.csail.mit.edu/papers/volume6/quinonero-candela05a/quinonero-candela05a.pdf.

[317] K. Eggensperger, M. Feurer, F. Hutter, J. Bergstra, J. Snoek, H. Hoos, and K. Leyton-Brown, "Towards an empirical foundation for assessing bayesian optimization of hyperparameters," In *NeurIPS Workshop on Bayesian Optimization in Theory and Practice*, Vol. 10, 2013, P. 3. [Online]. Available: https://www.cs.ubc.ca/~hoos/Publ/EggEtAl13.pdf.

[318] L. Breiman, "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)," *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, vol. 16, no. 3, pp. 199–231, Aug. 2001. DOI: 10.1214/ss/1009213726.

[319] R. B. Gramacy, M. Taddy, and S. M. Wild, "Variable selection and sensitivity analysis using dynamic trees, with an application to computer code performance tuning," *Annals of Applied Statistics*, vol. 7, no. 1, pp. 51–80, Mar. 2013. DOI: 10.1214/12-AOAS590.

[320] F. Hutter, H. Hoos, and K. Leyton-Brown, "An efficient approach for assessing hyperparameter importance," E. P. Xing and T. Jebara, Eds., Vol. 32, PMLR, Jun. 2014, Pp. 754–762. [Online]. Available: http://proceedings.mlr.press/v32/hutter14.pdf.

[321] P. Dufossé and C. Touré, "Benchmarking MO-CMA-ES and COMO-CMA-ES on the bi-objective bbob-biobj testbed," In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 2019, Pp. 1920–1927. DOI: doi:10.1145/3319619.3326892.

[322] L. Faury, C. Calauzènes, and O. Fercoq, "Benchmarking GNN-CMA-ES on the BBOB noiseless testbed," In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 2019, Pp. 1928–1936. DOI: 10.1145/3319619.3326856.

[323] B. Bodner, "Benchmarking the ATM algorithm on the BBOB 2009 noiseless function testbed,"

In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 2019,
Pp. 1897–1904.
DOI: 10.1145/3319619.3326802.

[324] C. Wehmeyer and F. Noé, "Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics,"
The Journal of Chemical Physics, vol. 148, no. 24, p. 241 703, Jun. 2018.
DOI: 10.1063/1.5011399.

[325] P. G. Bolhuis, C. Dellago, and D. Chandler, "Reaction coordinates of biomolecular isomerization,"
Proceedings of the National Academy of Sciences, vol. 97, no. 11, pp. 5877–5882, 2000.
DOI: 10.1073/pnas.100127697.

[326] F. Nüske, H. Wu, J.-H. Prinz, C. Wehmeyer, C. Clementi, and F. Noé, "Markov state models from short non-equilibrium simulations—Analysis and correction of estimation bias,"
The Journal of Chemical Physics, vol. 146, no. 9, p. 094 104, Mar. 2017.
DOI: 10.1063/1.4976518.

[327] F. Nüske, L. Boninsegna, and C. Clementi, "Coarse-graining molecular systems by spectral matching,"
The Journal of Chemical Physics, vol. 151, no. 4, p. 044 116, Jul. 2019.
DOI: 10.1063/1.5100131.

[328] J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. de Fabritiis, F. Noé, and C. Clementi, "Machine Learning of Coarse-Grained Molecular Dynamics Force Fields,"
ACS Central Science, vol. 5, no. 5, pp. 755–767, May 2019.
DOI: 10.1021/acscentsci.8b00913.

[329] S.-H. Li, C.-X. Dong, L. Zhang, and L. Wang, "Neural Canonical Transformation with Symplectic Flows,"
Physical Review X, vol. 10, no. 2, p. 021 020, Apr. 2020.
DOI: 10.1103/PhysRevX.10.021020.

[330] Y. B. Varolgüneş, T. Bereau, and J. F. Rudzinski, "Interpretable embeddings from molecular simulations using Gaussian mixture variational autoencoders,"
Machine Learning: Science and Technology, vol. 1, no. 1, p. 015 012, Mar. 2020.
DOI: 10.1088/2632-2153/ab80b7.

[331] F. Nüske, P. Koltai, L. Boninsegna, and C. Clementi, "Spectral Properties of Effective Dynamics from Conditional Expectations,"
Entropy, vol. 23, no. 2, p. 134, Feb. 2021.
DOI: 10.3390/e23020134.

[332] R. Sechi, A. Sikorski, and M. Weber, "Estimation of the Koopman Generator by Newton's Extrapolation,"
Multiscale Modeling & Simulation, vol. 19, no. 2, pp. 758–774, Jan. 2021.
DOI: 10.1137/20M1333006.

[333] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water,"
The Journal of chemical physics, vol. 79, no. 2, pp. 926–935, 1983.
DOI: 10.1063/1.445869.

[334] M. J. Harvey, G. Giupponi, and G. D. Fabritiis, "ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale,"
Journal of Chemical Theory and Computation, vol. 5, no. 6, pp. 1632–1639, Jun. 2009.
DOI: 10.1021/ct9000685.

[335] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99SB protein force field,"
Proteins: Structure, Function, and Bioinformatics, vol. 78, no. 8, pp. 1950–1958, 2010.
DOI: 10.1002/prot.22711.

[336] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems,"
The Journal of Chemical Physics, vol. 98, no. 12, pp. 10 089–10 092, Jun. 1993.
DOI: 10.1063/1.464397.

[337] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, "MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories,"
Biophysical Journal, vol. 109, no. 8, pp. 1528–1532, 2015.
DOI: 10.1016/j.bpj.2015.08.015.

[338] S. van der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy Array: A Structure for Efficient Numerical Computation,"
Computing in Science Engineering, vol. 13, no. 2, pp. 22–30, 2011.

[339] W. McKinney, "Pandas: A foundational Python library for data analysis and statistics,"
Python for High Performance and Scientific Computing, vol. 14, 2011.

[340] J. D. Hunter, "Matplotlib: A 2D graphics environment,"
Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.
DOI: 10.1109/MCSE.2007.55.

[341] M. Waskom, O. Botvinnik, J. Ostblom, M. Gelbart, S. Lukauskas, P. Hobson, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, Julian de Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, C. Swain, A. Miles, T. Brunner, D. O'Kane, T. Yarkoni, M. L. Williams, C. Evans, C. Fitzgerald, and Brian, *Mwaskom/seaborn: V0.10.1 (April 2020)*,
Zenodo,
Apr. 2020.
DOI: 10.5281/zenodo.3767070.

[342] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing, and J. development team, "Jupyter Notebooks ? a publishing format for reproducible computational workflows,"
In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*,
F. Loizides and B. Scmidt, Eds.,
IOS Press, 2016,
Pp. 87–90.

[343] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in Python using PyMC3,"
PeerJ Computer Science, vol. 2, e55, Apr. 2016.
DOI: 10.7717/peerj-cs.55.

[344] D. Duvenaud,
"Automatic model construction with Gaussian processes,"
Thesis, University of Cambridge, Nov. 2014.
DOI: 10.17863/CAM.14087.

[345] N. G. Polson and J. G. Scott, "On the Half-Cauchy Prior for a Global Scale Parameter,"
Bayesian Analysis, vol. 7, no. 4, pp. 887–902, Dec. 2012.
DOI: 10.1214/12-BA730.

[346] A. Daly, T. Dekker, and S. Hess, "Dummy coding vs effects coding for categorical variables: Clarifications and extensions,"
Journal of Choice Modelling, vol. 21, pp. 36–41, Dec. 2016.
DOI: 10.1016/j.jocm.2016.09.005.

[347] GPy Authors,
GPy: A Gaussian process framework in python,
2012.
[Online]. Available: https://github.com/SheffieldML/GPy.

[348] R. M. Neal,
"Regression and classification using Gaussian process priors,"
In *Bayesian statistics 6: Proceedings of the Sixth Valencia International Meeting*,

J. Bernardo, J. Berger, A. Dawid, A. Smith, and A. P. Dawid, Eds.,
Vol. 6, Oxford University Press, 1998,
Pp. 475–501.

[349] F. Harrel,
Regression Modeling Strategies,
Second edition.
Springer International Publishing, 2015.

[350] D. K. Duvenaud, H. Nickisch, and C. E. Rasmussen,
"Additive gaussian processes,"
In *Advances in Neural Information Processing Systems 24*,
J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds.,
Curran Associates, Inc., 2011,
Pp. 226–234.
[Online]. Available: `http://papers.nips.cc/paper/4221-additive-gaussian-processes.pdf`.

[351] G. R. Bowman, X. Huang, and V. S. Pande, "Network models for molecular kinetics and their initial applications to human health,"
Cell Research, vol. 20, no. 6, pp. 622–630, Jun. 2010.
DOI: 10.1038/cr.2010.57.

[352] K. Henzler-Wildman and D. Kern, "Dynamic personalities of proteins,"
Nature, vol. 450, no. 7172, pp. 964–972, Dec. 2007.
DOI: 10.1038/nature06522.

[353] A. Jain and G. Stock, "Hierarchical Folding Free Energy Landscape of HP35 Revealed by Most Probable Path Clustering,"
The Journal of Physical Chemistry B, vol. 118, no. 28, pp. 7750–7760, Jul. 2014.
DOI: 10.1021/jp410398a.

[354] K. G. Wilson, "The renormalization group and critical phenomena,"
Reviews of Modern Physics, vol. 55, no. 3, pp. 583–600, Jul. 1983.
DOI: 10.1103/RevModPhys.55.583.

[355] A. Kells, Z. É. Mihálka, A. Annibale, and E. Rosta, "Mean first passage times in variational coarse graining using Markov state models,"
The Journal of Chemical Physics, vol. 150, no. 13, p. 134 107, Apr. 2019.
DOI: 10.1063/1.5083924.

[356] G. S. Jas and K. Kuczera, "Helix–Coil Transition Courses Through Multiple Pathways and Intermediates: Fast Kinetic Measurements and Dimensionality Reduction,"
The Journal of Physical Chemistry B, vol. 122, no. 48, pp. 10 806–10 816, Dec. 2018.
DOI: 10.1021/acs.jpcb.8b07924.

[357] G. Claeskens and N. L. Hjort,
Model Selection and Model Averaging.
Cambridge University Press, 2008.
DOI: 10.1017/CBO9780511790485.

[358] D. J. MacKay,
Information Theory, Inference and Learning Algorithms.
Cambridge university press, 2003.

[359] G. J. McLachlan and K. E. Basford,
Mixture Models: Inference and Applications to Clustering.
Marcel Dekker, Inc., 1988.

[360] D. A. Reynolds, "Gaussian mixture models,"
Encyclopedia of biometrics, vol. 741, 2009.
[Online]. Available: http://leap.ee.iisc.ac.in/sriram/teaching/MLSP_16/refs/GMM_Tutorial_Reynolds.pdf.

[361] R. J. Hathaway, "Another interpretation of the EM algorithm for mixture distributions,"
Statistics & probability letters, vol. 4, no. 2, pp. 53–56, 1986.
DOI: 10.1016/0167-7152(86)90016-7.

[362] P. Latouche, E. Birmelé, and C. Ambroise,
"Bayesian Methods for Graph Clustering,"
In *Advances in Data Analysis, Data Handling and Business Intelligence*,
A. Fink, B. Lausen, W. Seidel, and A. Ultsch, Eds.,
Springer, 2010,
Pp. 229–239.
DOI: 10.1007/978-3-642-01044-6_21.

[363] C. Fraley and A. E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis,"
The computer journal, vol. 41, no. 8, pp. 578–588, 1998.
DOI: 10.1093/comjnl/41.8.578.

[364] S. Kullback and R. A. Leibler, "On Information and Sufficiency,"
Annals of Mathematical Statistics, vol. 22, no. 1, pp. 79–86, Mar. 1951.
DOI: 10.1214/aoms/1177729694.

[365] B. D. Ripley,
Pattern Recognition and Neural Networks.
Cambridge University Press, 1996.
DOI: 10.1017/CBO9780511812651.

[366] J. Bernardo and A. Smith,

Bayesian Theory.
Wiley, 2007.

[367] E. Gassiat, "Likelihood ratio inequalities with applications to various mixtures,"
Annales de l'Institut Henri Poincare (B) Probability and Statistics, vol. 38, no. 6, pp. 897–906, Nov. 2002.
DOI: 10.1016/S0246-0203(02)01125-1.

[368] G. Celeux and G. Soromenho, "An entropy criterion for assessing the number of clusters in a mixture model,"
Journal of Classification, vol. 13, no. 2, pp. 195–212, Sep. 1996.
DOI: 10.1007/BF01246098.

[369] G. Soromenho, "Comparing approaches for testing the number of components in a finite mixture model,"
Computational Statistics, vol. 9, no. 1, pp. 65–78, 1994.

[370] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins,"
The Journal of Physical Chemistry B, vol. 102, no. 18, pp. 3586–3616, 1998, PMID: 24889800.
DOI: 10.1021/jp973084f.

[371] L. Masgrau, K. E. Ranaghan, N. S. Scrutton, A. J. Mulholland, and M. J. Sutcliffe, "Tunneling and classical paths for proton transfer in an enzyme reaction dominated by tunneling: Oxidation of tryptamine by aromatic amine dehydrogenase,"
Journal of Physical Chemistry B, vol. 111, no. 11, pp. 3032–3047, 2007.
DOI: 10.1021/jp067898k.

[372] K. E. Ranaghan, W. G. Morris, L. Masgrau, K. Senthilkumar, L. O. Johannissen, N. S. Scrutton, J. N. Harvey, F. R. Manby, and A. J. Mulholland, "Ab Initio QM/MM Modeling of the Rate-Limiting Proton Transfer Step in the Deamination of Tryptamine by Aromatic Amine Dehydrogenase,"
The Journal of Physical Chemistry B, vol. 121, no. 42, pp. 9785–9798, Oct. 2017.
DOI: 10.1021/acs.jpcb.7b06892.

[373] A. Y. Chistoserdov, "Cloning, sequencing and mutagenesis of the genes for aromatic amine dehydrogenase from Alcaligenes faecalis and evolution of amine dehydrogenases. The GenBank accession number for the aau gene cluster from Alcaligenes faecalis is AF302652,"
Microbiology, vol. 147, no. 8, pp. 2195–2202, 2001.
DOI: 10.1099/00221287-147-8-2195.

[374] S. Govindaraj, E. Eisenstein, L. H. Jones, J. Sanders-Loehr, A. Y. Chistoserdov, V. L. Davidson, and S. L. Edwards, "Aromatic amine dehydrogenase, a second tryptophan tryptophylquinone enzyme.,"
Journal of Bacteriology, vol. 176, no. 10, pp. 2922–2929, May 1994.
DOI: 10.1128/jb.176.10.2922-2929.1994.

[375] M. Iwaki, T. Yagi, K. Horiike, Y. Saeki, T. Ushijima, and M. Nozaki, "Crystallization and properties of aromatic amine dehydrogenase from Pseudomonas sp,"
Archives of Biochemistry and Biophysics, vol. 220, no. 1, pp. 253–262, Jan. 1983.
DOI: 10.1016/0003-9861(83)90408-3.

[376] A. Roujeinikova, P. Hothi, L. Masgrau, M. J. Sutcliffe, N. S. Scrutton, and D. Leys, "New Insights into the Reductive Half-reaction Mechanism of Aromatic Amine Dehydrogenase Revealed by Reaction with Carbinolamine Substrates,"
Journal of Biological Chemistry, vol. 282, no. 33, pp. 23 766–23 777, Aug. 2007.
DOI: 10.1074/jbc.M700677200.

[377] N. Sukumar, Z.-w. Chen, D. Ferrari, A. Merli, G. L. Rossi, H. D. Bellamy, A. Chistoserdov, V. L. Davidson, and F. S. Mathews, "Crystal Structure of an Electron Transfer Complex between Aromatic Amine Dehydrogenase and Azurin from Alcaligenes faecalis,"
Biochemistry, vol. 45, no. 45, pp. 13 500–13 510, Nov. 2006.
DOI: 10.1021/bi0612972.

[378] P. Hothi, A. Roujeinikova, K. A. Khadra, M. Lee, P. Cullis, D. Leys, and N. S. Scrutton, "Isotope Effects Reveal That Para-Substituted Benzylamines Are Poor Reactivity Probes of the Quinoprotein Mechanism for Aromatic Amine Dehydrogenase,"
Biochemistry, vol. 46, no. 32, pp. 9250–9259, Aug. 2007.
DOI: 10.1021/bi7007239.

[379] Y.-L. Hyun and V. L. Davidson, "Mechanistic Studies of Aromatic Amine Dehydrogenase, a Tryptophan Tryptophylquinone Enzyme,"
Biochemistry, vol. 34, no. 3, pp. 816–823, Jan. 1995.
DOI: 10.1021/bi00003a015.

[380] F. Claeyssens, J. N. Harvey, F. R. Manby, R. A. Mata, A. J. Mulholland, K. E. Ranaghan, M. Schütz, S. Thiel, W. Thiel, and H.-J. Werner, "High-Accuracy Computation of Reaction Barriers in Enzymes,"
Angewandte Chemie, vol. 118, no. 41, pp. 7010–7013, 2006.
DOI: 10.1002/ange.200602711.

[381] M. Kazemi, F. Himo, and J. Åqvist, "Enzyme catalysis by entropy without Circe effect,"
Proceedings of the National Academy of Sciences, vol. 113, no. 9, pp. 2406–2411, Mar. 2016.

[382] H. Bisswanger,
Enzyme Kinetics: Principles and Methods.

John Wiley & Sons, 2017.

[383] Y.-L. Hyun and V. L. Davidson, "Unusually large isotope effect for the reaction of aromatic amine dehydrogenase. A common feature of Quinoproteins?"
Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology, vol. 1251, no. 2, pp. 198–200, Sep. 1995.
DOI: 10.1016/0167-4838(95)00117-D.

[384] R. P. Bell,
The Tunnel Effect in Chemistry.
Springer US, 1980.
DOI: 10.1007/978-1-4899-2891-7.

[385] D. Antoniou and S. D. Schwartz, "Internal Enzyme Motions as a Source of Catalytic Activity: Rate-Promoting Vibrations and Hydrogen Tunneling,"
The Journal of Physical Chemistry B, vol. 105, no. 23, pp. 5553–5558, Jun. 2001.
DOI: 10.1021/jp004547b.

[386] Y. Cha, C. J. Murray, and J. P. Klinman, "Hydrogen tunneling in enzyme reactions,"
Science, vol. 243, no. 4896, pp. 1325–1330, Mar. 1989.
DOI: 10.1126/science.2646716.

[387] A. Kohen, "Role of Dynamics in Enzyme Catalysis: Substantial versus Semantic Controversies,"
Accounts of Chemical Research, vol. 48, no. 2, pp. 466–473, Feb. 2015.
DOI: 10.1021/ar500322s.

[388] S. Hay and N. S. Scrutton, "Good vibrations in enzyme-catalysed reactions,"
Nature Chemistry, vol. 4, no. 3, pp. 161–168, 2012.
DOI: 10.1038/nchem.1223.

[389] A. Fersht,
Structure and Mechanism in Protein Science. A Guide to Enzyme Catalysis and Protein Folding.
W. H. Freeman and co., 1999.

[390] S. C. L. Kamerlin and A. Warshel, "At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis?"
Proteins, vol. 78, no. 6, pp. 1339–1375, May 2010.
DOI: 10.1002/prot.22654.

[391] U. Doshi and D. Hamelberg, "The dilemma of conformational dynamics in enzyme catalysis: Perspectives from theory and experiment,"
Advances in Experimental Medicine and Biology, vol. 805, pp. 221–243, 2014.
DOI: 10.1007/978-3-319-02970-2_10.

[392] A. Warshel and R. P. Bora, "Perspective: Defining and quantifying the role of dynamics in enzyme catalysis,"

The Journal of Chemical Physics, vol. 144, no. 18, p. 180 901, 2016.
DOI: 10.1063/1.4947037.

[393] N. Kanaan, M. Roca, I. Tuñón, S. Martí, and V. Moliner, "Theoretical study of the temperature dependence of dynamic effects in thymidylate synthase,"
Physical Chemistry Chemical Physics, vol. 12, no. 37, pp. 11 657–11 664, 2010.
DOI: 10.1039/C003799K.

[394] M. Roca, V. Moliner, I. Tuñón, and J. T. Hynes, "Coupling between Protein and Reaction Dynamics in Enzymatic Processes: Application of Grote-Hynes Theory to Catechol O-Methyltransferase,"
Journal of the American Chemical Society, vol. 128, no. 18, pp. 6186–6193, May 2006.
DOI: 10.1021/ja058826u.

[395] J. Huang and A. D. Mackerell, "CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data,"
Journal of Computational Chemistry, vol. 34, no. 25, pp. 2135–2145, 2013.
DOI: 10.1002/jcc.23354.

[396] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, "CHARMM: The biomolecular simulation program,"
Journal of Computational Chemistry, vol. 30, no. 10, pp. 1545–1614, 2009.
DOI: 10.1002/jcc.21287.

[397] H. Grubmüller and V. Groll,
Solvate.
[Online]. Available: https://www.mpibpc.mpg.de/grubmueller/solvate.

[398] W. Humphrey, A. Dalke, and K. Schulten, "VMD – Visual Molecular Dynamics,"
Journal of Molecular Graphics, vol. 14, pp. 33–38, 1996.
DOI: 10.1016/0263-7855(96)00018-5.

[399] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande, "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics,"
PLOS Computational Biology, vol. 13, no. 7, e1005659, Jul. 2017.
DOI: 10.1371/journal.pcbi.1005659.

[400] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes,"
Journal of Computational Physics, vol. 23, no. 3, pp. 327–341, Mar. 1977.
DOI: 10.1016/0021-9991(77)90098-5.

[401] D. A. Case, R. M. Betz, D. S. Cerutti, T. E. Cheatham III, T. A. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K. M. Merz, G. Monard, H. Nguyen, H. T. Nguyen, I. Omelyan, A. Onufriev, D. R. Roe, A. Roitberg, C. Sagui, C. L. Simmerling, W. M. Botello-Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, L. Xiao, and P. A. Kollman,
AMBER 2016,
2016.

[402] K. Swersky, D. Duvenaud, J. Snoek, F. Hutter, and M. A. Osborne, "Raiders of the Lost Architecture: Kernels for Bayesian Optimization in Conditional Parameter Spaces,"
Sep. 2014.
arXiv: 1409.4011 [stat].

[403] S. Doerr and G. De Fabritiis, "On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations,"
Journal of Chemical Theory and Computation, vol. 10, no. 5, pp. 2064–2069, May 2014.
DOI: 10.1021/ct400919u.

[404] B. C. Hancock and G. Zografi, "Characteristics and significance of the amorphous state in pharmaceutical systems,"
Journal of Pharmaceutical Sciences, vol. 86, no. 1, pp. 1–12, Jan. 1997.
DOI: 10.1021/js9601896.

[405] C. W. Fox and S. J. Roberts, "A tutorial on variational Bayesian inference,"
Artificial Intelligence Review, vol. 38, no. 2, pp. 85–95, Aug. 2012.
DOI: 10.1007/s10462-011-9236-8.

[406] K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande, "MSMBuilder2: Modeling conformational dynamics on the picosecond to millisecond scale,"
Journal of Chemical Theory and Computation, vol. 7, no. 10, pp. 3412–3419, 2011.
DOI: 10.1021/ct200463m.

[407] C. Gu, H.-W. Chang, L. Maibaum, V. S. Pande, G. E. Carlsson, and L. J. Guibas, "Building Markov state models with solvent dynamics,"
BMC Bioinformatics, vol. 14, no. 2, S8, Jan. 2013.
DOI: 10.1186/1471-2105-14-S2-S8.

[408] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine Learning for Molecular Simulation,"
Annual Review of Physical Chemistry, vol. 71, no. 1, pp. 361–390, 2020.
DOI: 10.1146/annurev-physchem-042018-052331.

[409] R. M. Power and J. P. Reid, "Probing the micro-rheological properties of aerosol particles using optical tweezers,"

Reports on Progress in Physics, vol. 77, no. 7, p. 074 601, Jul. 2014.
DOI: 10.1088/0034-4885/77/7/074601.

[410] T. C. Preston and J. P. Reid, "Accurate and efficient determination of the radius, refractive index, and dispersion of weakly absorbing spherical particle using whispering gallery modes," *Journal of the Optical Society of America B*, vol. 30, no. 8, pp. 2113–2122, Aug. 2013.
DOI: 10.1364/JOSAB.30.002113.

[411] J. P. Reid, H. Meresman, L. Mitchem, and R. Symes†, "Spectroscopic studies of the size and composition of single aerosol droplets,"
International Reviews in Physical Chemistry, vol. 26, no. 1, pp. 139–192, Jan. 2007.
DOI: 10.1080/01442350601081899.

[412] J. P. Reid and L. Mitchem, "Laser probing of single-aerosol droplet dynamics,"
Annual Review of Physical Chemistry, vol. 57, no. 1, pp. 245–271, Apr. 2006.
DOI: 10.1146/annurev.physchem.57.032905.104621.

[413] T. Chen, A. Fowler, and M. Toner, "Literature Review: Supplemented Phase Diagram of the Trehalose–Water Binary Mixture,"
Cryobiology, vol. 40, no. 3, pp. 277–282, May 2000.
DOI: 10.1006/cryo.2000.2244.

[414] K. D. Foster, J. E. Bronlund, and A. H. J. (Paterson, "Glass transition related cohesion of amorphous sugar powders,"
Journal of Food Engineering, vol. 77, no. 4, pp. 997–1006, Dec. 2006.
DOI: 10.1016/j.jfoodeng.2005.08.028.

[415] J. W. Lu, A. M. J. Rickards, J. S. Walker, K. J. Knox, R. E. H. Miles, J. P. Reid, and R. Signorell, "Timescales of water transport in viscous aerosol: Measurements on sub-micron particles and dependence on conditioning history,"
Physical Chemistry Chemical Physics, vol. 16, no. 21, pp. 9819–9830, May 2014.
DOI: 10.1039/C3CP54233E.

[416] S. O'Meara, D. O. Topping, and G. McFiggans, "The rate of equilibration of viscous aerosol particles,"
Atmospheric Chemistry and Physics, vol. 16, no. 8, pp. 5299–5313, Apr. 2016.
DOI: 10.5194/acp-16-5299-2016.

[417] J. L. F. Abascal and C. Vega, "A general purpose model for the condensed phases of water: TIP4P/2005,"
The Journal of Chemical Physics, vol. 123, no. 23, p. 234 505, Dec. 2005.
DOI: 10.1063/1.2121687.

[418] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. V. Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6,"
Journal of Computational Chemistry, vol. 25, no. 13, pp. 1656–1676, 2004.
DOI: 10.1002/jcc.20090.

[419] K. B. Koziara, M. Stroet, A. K. Malde, and A. E. Mark, "Testing and validation of the Automated Topology Builder (ATB) version 2.0: Prediction of hydration free enthalpies,"
Journal of Computer-Aided Molecular Design, vol. 28, no. 3, pp. 221–233, Mar. 2014.
DOI: 10.1007/s10822-014-9713-7.

[420] A. K. Malde, L. Zuo, M. Breeze, M. Stroet, D. Poger, P. C. Nair, C. Oostenbrink, and A. E. Mark, "An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0,"
Journal of Chemical Theory and Computation, vol. 7, no. 12, pp. 4026–4037, Dec. 2011.
DOI: 10.1021/ct200196m.

[421] N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark, and W. F. van Gunsteren, "Definition and testing of the GROMOS force-field versions 54A7 and 54B7,"
European Biophysics Journal, vol. 40, no. 7, p. 843, Apr. 2011.
DOI: 10.1007/s00249-011-0700-9.

[422] L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez, "Packmol: A package for building initial configurations for molecular dynamics simulations,"
Journal of computational chemistry, vol. 30, no. 13, pp. 2157–2164, 2009.
DOI: 10.1002/jcc.21224.

[423] B. Hess, H. Bekker, H. J. Berendsen, and J. G. Fraaije, "LINCS: A linear constraint solver for molecular simulations,"
Journal of computational chemistry, vol. 18, no. 12, pp. 1463–1472, 1997.
DOI: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

[424] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, "A smooth particle mesh Ewald method,"
The Journal of Chemical Physics, vol. 103, no. 19, pp. 8577–8593, Nov. 1995.
DOI: 10.1063/1.470117.

[425] S. Páll and B. Hess, "A flexible algorithm for calculating pair interactions on SIMD architectures,"
Computer Physics Communications, vol. 184, no. 12, pp. 2641–2650, Dec. 2013.
DOI: 10.1016/j.cpc.2013.06.003.

[426] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method,"
Journal of Applied Physics, vol. 52, no. 12, pp. 7182–7190, Dec. 1981.
DOI: 10.1063/1.328693.

[427] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers,"
SoftwareX, vol. 1-2, pp. 19–25, Sep. 2015.
DOI: 10.1016/j.softx.2015.06.001.

[428] X. Michalet, "Mean square displacement analysis of single-particle trajectories with localization error: Brownian motion in an isotropic medium,"
Physical Review E, vol. 82, no. 4, p. 041 914, Oct. 2010.
DOI: 10.1103/PhysRevE.82.041914.

[429] N. Alcázar-Cano and R. Delgado-Buscalioni, "A general phenomenological relation for the subdiffusive exponent of anomalous diffusion in disordered media,"
Soft Matter, vol. 14, no. 48, pp. 9937–9949, Dec. 2018.
DOI: 10.1039/C8SM01961D.

[430] M. L. Meste, D. Champion, G. Roudaut, G. Blond, and D. Simatos, "Glass Transition and Food Technology: A Critical Appraisal,"
Journal of Food Science, vol. 67, no. 7, pp. 2444–2458, 2002.
DOI: 10.1111/j.1365-2621.2002.tb08758.x.

[431] B. Zobrist, V. Soonsin, B. P. Luo, U. K. Krieger, C. Marcolli, T. Peter, and T. Koop, "Ultra-slow water diffusion in aqueous sucrose glasses,"
Physical Chemistry Chemical Physics, vol. 13, no. 8, pp. 3514–3526, 2011.
DOI: 10.1039/C0CP01273D.

[432] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features,"
Biopolymers, vol. 22, no. 12, pp. 2577–2637, 1983.
DOI: 10.1002/bip.360221211.